CROSS-LAYER SCHEDULING AND TRANSMISSION STRATEGIES FOR ENERGY-CONSTRAINED WIRELESS NETWORKS

HOANG ANH TUAN

(B.Eng. Hons., Uni. of Sydney)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2005

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest gratitude to my supervisor, Dr. Mehul Motani, for so much time and effort he has spent on guiding me through every state of this thesis. Dr. Motani has been a continuous source of ideas, encouragement, motivation, and support for me. While always being available to help, he also gave me room to independently explore different directions; this clearly made the research work much more enjoyable for me. I feel truly fortunate to have been working under Dr. Motani's guidance.

Next, I would like to thank friends in my research group, including Vineet Srivastava, Lawrence Ong, Kok Kiong Yap, and Hon Fah Chong, for many interesting research discussions that not only improved my thesis, but also broadened my knowledge. In particular, I want to give a special thanks to Vineet for his true friendship.

I want to thank my parents, brother Dung, and sister Phuong, for their love and care. Especially, I am indebted to my parents for so many sacrifices they have made for me. My parents have made me who I am.

Finally, I am most grateful to my wife, Cuc, and my *'little Pinocchio'*, Minh, for sharing this journey with me. It was tough to have a hubby who was usually lost in models and equations; it was also tough to have a daddy who was usually busy at weekends. Despite all this, Cuc and Minh have always been the ones I turned to for love and encouragement. I thank them for being a part of my life.

TABLE OF CONTENTS

SUMMARY

Recently, cross-layer design has been identified as a promising approach which achieves good performance for energy-constrained wireless networks. In general, cross-layer design refers to the methodology in which multiple layers in the communication protocol stack are designed in an integrated manner, with the intra-layer and inter-layer dynamics being taken into account. In this thesis, we study cross-layer scheduling and transmission strategies that provide good system performance, in terms of throughput, while conserving nodes' energy.

First, we consider a cross-layer adaptive transmission problem for single-user systems with stochastic data arrivals, finite-length buffer operating over a time-varying wireless channel. The objective is to adapt the transmit power and rate according to the buffer and channel conditions so that the system throughput is maximized, subject to an average transmit power constraint. We demonstrate that this problem can be solved by reformulating it as a Markov decision process. We then identify an important structural characteristic of the throughput optimal policy, which is in sharp contrast to the structure of policies that achieve capacity of fading channels. We also consider the adaptive transmission problem when only a partial observation of the buffer or channel states is available.

Next, we consider a multiple-access scenario in which multiple users share a single channel to transmit data to a center node. There are two control decisions to be made in each time slot, i.e., a scheduling decision which assigns the channel to one of the users, and a transmission decision which sets the transmit power and rate. All scheduling/transmission policies employed must satisfy the average transmit power constraint of each node. We first look at

the problem of finding the optimal cross-layer adaptive scheduling/transmission policy which adapts to the buffer and channel conditions of all users so that the total system throughput is maximized. We then use the performance of this optimal policy as a benchmark to assess the performance of simpler adaptive scheduling/transmission schemes which also adapt to the buffer and channel conditions. This allows us to draw some useful guidelines for controlling energy-constrained multiple-access systems.

Finally, we study a problem of combining scheduling, transmission, and data compression to conserve energy in a spatially correlated cluster-based sensor networks. Since wireless transmission is inherently broadcast, when one sensor node transmits data to the cluster head, other nodes in its coverage area can receive the transmitted data. When data collected by different sensors are correlated, each sensor can utilize the data it overhears from others' transmissions to compress its own data and conserve energy in its own transmissions. Based on this observation, we formulate a problem in which sensors in each cluster are scheduled to transmit so that they can collaborate in joint source compression in order to maximize the network lifetime. We show that this lifetime optimization problem can be solved by a sequence of linear programming problems. We also develop a heuristic scheme which has low complexity and achieves near optimal performance.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AWGN	Additive Wide Gaussian Noise
BER	Bit Error Rate
BIP	Broadcast Incremental Power
CBC	Collaborative Broadcasting and Compression
CSI	Channel State Information
FIFO	First In First Out
FSMC	Finite State Markov Channel
LP	Linear Programming
LVO	Lifetime Vector Optimization
MAC	Media Access Control (layer)
MDP	Markov Decision Process
MIC	Minimum Immediate Cost
MLS	Most Likely State
M-LWDF	Modified Largest Weighted Delay First
M-LWWF	Modified Largest Weighted Work First
MQAM	M-ary Quadrature Amplitude Modulation
PER	Packet Error Rate
PHY	Physical (layer)
POMDP	Partially Observable Markov Decision Process
SSI	System State Information
WLAN	Wireless Local Area Network
WSN	Wireless Sensor Network

CHAPTER 1

INTRODUCTION

Many modern and future wireless networks comprise nodes that operate based on small and energy-limited batteries. Examples of such networks include mobile cellular systems, wireless local area networks, wireless ad hoc networks, and wireless sensor networks. In these energy-constrained wireless networks, a fundamental design challenge is to achieve good system performance while conserving nodes' energy. We address this challenge by studying different energy-efficient scheduling and transmission strategies for wireless networks. In doing so, we adopt the *cross-layer design* approach, which designs and controls the operations of different layers of the network architecture in an integrated fashion. This is in contrast to the popular layered design approach, that has been widely followed in designing wired computer networks. This chapter gives the background information of our research, the specific problems we study, and the main contributions we have made.

1.1 Energy-constrained Wireless Networks

Based on their architecture, wireless networks can be classified into two main categories, i.e., *infrastructure-based* wireless networks and *infrastructure-less* wireless networks. In both categories, there are wireless nodes that operate with highly limited power and energy sources.

1.1.1 Infrastructure-based Wireless Networks

Infrastructure-based wireless networks are set up based on some preexisting network backbones. These backbones comprise wired or microwave links capable of carrying data at high speeds. Thus, in terms of the system connectivity, the main function of an infrastructure-based wireless network is to provide the wireless extension from a backbone to wireless devices.

Examples of infrastructure-based wireless networks are cellular mobile communication systems. In such a system, the geographical area is divided into subareas called cells. Within each cell placed a base station that directly communicates with all mobile terminals locating in the cell over the wireless medium. Base stations are linked by a backbone network which is connected to the public switch telephone network (PSTN) or the Internet through a number of gateways. In addition to voice services, modern and future cellular systems also support data and multi-media applications.

Other examples of infrastructure-based wireless networks are wireless local area networks (WLANs). Today, WLANs following the IEEE 802.11 standard ([CWKS97]) are becoming more and more popular. An WLAN consists of a number of access points that are wired to the Internet backbone. Wireless devices such as laptops and personal digital assistants (PDAs) communicate with a nearby access point using wireless transmission. Note that apart from this star-topology, peer-to-peer architecture is also supported by the IEEE 802.11 standard.

It should be noted that, as base stations and access points are connected to some network backbone with stable power supplies, energy constraint is usually not a critical design issue for the *downlink*, i.e., the link from base stations or access points toward wireless terminals. On the other hand, wireless terminals such as mobile phones, PDAs, and laptops are small in size and can only be equipped with limited batteries. Moreover, the users of these devices can access wireless services while on the move, making batteries recharging and/or replacing undesirable. As a result, power and energy constraints must be taken care of in the design of the *uplink*, i.e., the link from wireless devices toward base stations or access points.

1.1.2 Infrastructure-less Wireless Networks

Infrastructure-less wireless networks are designed to be deployed without the support of any preexisting network infrastructure. With respect to infrastructure-based networks, they have the advantages of shorter deployment time, flexibility in network architecture, and robust to single-point failures [GW02]. Examples of infrastructure-less wireless networks are wireless ad hoc networks and wireless sensor networks.

In wireless ad hoc networks, connectivity is built upon peer-to-peer communication between nodes. When two wireless nodes are far apart so that no direct communication is possible, connectivity can be provided by multihop routing. This leads to the fact that, depending on the network topology and routing decisions, nodes may have to act as both data hosts and routers. When this is the case, the energy and power constraints of a node affect not only its own performance, but also the performance of other nodes that utilize it as router/relayer. As a result, power and energy conservation is a critical design criterion for wireless ad hoc networks.

A wireless sensor network (WSN) consists of a large number of low-cost, low-power, and tiny sensors. These sensors are capable of collecting statistics from the environment, processing collected information, and communicating data toward some command centers using wireless transmission. In the literature, WSNs are sometimes regarded as a class of wireless ad hoc networks [GW02]. However, it can be argued that, with respect to general wireless ad hoc networks, WSNs deserve a separate treatment due to the following reasons. First of all, a WSN can be much denser compared to a typical wireless ad hoc network. Due to the short distance between sensor nodes, the energy consumed in data transmission is greatly reduced. In fact, this energy consumption is comparable to the energy consumed in the processors and electronic circuits of sensor nodes. This means that for each sensor node, energy consumed in processing, receiving, and transmitting must all be taken into consideration. Secondly, nodes of WSNs can be much smaller than those of a typical wireless ad hoc network. Typical wireless ad hoc networks comprise laptops, PDAs, and other handheld devices. On the other hand, WSNs are envisaged to consist of nodes as small as a dust [KKP99]. This implies that sensor nodes are much more energy-constrained and prone to failure. Finally, a very special characteristic of WSNs is that data collected by sensors can be correlated. This is fundamentally different from the assumption of independent flows in the design of wireless ad hoc networks.

Before moving on, it is important to note that, even though wireless ad hoc networks and WSNs are infrastructure-less, their architectures need not be totally flat [GW02]. In particular, a hierarchical structure can be set up to assist data delivery. For example, in an wireless ad hoc network, some nodes can be elected to act as base stations or to form some network backbone to improve network reliability and capacity [Haa00, BTD01]. Similarly, sensor networks can be organized in to clusters, which each cluster being controlled by a cluster head [HCB00, HM05a].

1.2 Design Approaches

Layered design has been regarded as a major factor behind the proliferation of wired data networks [KK05]. However, for energy-constrained wireless networks, there are strong motivations for a more flexible design methodology, called *cross-layer design*, which can adapt and take advantage of various characteristics of the wireless medium [GW02, SRK03, KK05]. We will discuss these design approaches next.

1.2.1 Layered Architectures and Layered Design

Layered design is based on some layered network architectures. In such an architecture, the network functions are divided into different layers of a protocol stack. Protocols are designed within each layer, in a manner independent to the internal operation of other layers.

Examples of layered architectures are the Open System Interconnection (OSI) reference model and the TCP/IP model of the Internet. The OSI model consists of seven layers, i.e., from bottom up, *physical, data link, network, transport, session, presentation*, and *application*, while the TCP/IP has four, i.e., *link, network, transport*, and *application.* In these architectures, each layer utilizes the functions of the layer right below it in order to provide services for the layer above. It is important to note that interactions between adjacent layers are based on relatively static interfaces. For example, in the OSI model, the task of the physical layer is to provide a constant bit stream for the data link layer. In turn, the data link layer is expected to provide the network layer with some constant packet transmission rate and packet loss probability.

It is evident that the layered design approach has been a cornerstone for the success of the wired data networks in general and the Internet in particular [KK05]. By dividing the network functions into separate layers, it breaks down the complex task of network design into a set of independent and more manageable problems. The layered approach also allows engineers to work on designing different layers in parallel, i.e, work in one layer can be carried out without worrying about the detailed operation inside other layers. An important long-term effect is that the layered design approach ensures that continuous innovations can happen within each layer. By this we mean that each layer can be continuously optimized, as long as this conforms with the specifications of the layered architecture, the newly optimized layer will work fine with the rest of the protocol stack.

1.2.2 Cross-layer Design

Despite the success of the layered design approach for wired data networks, there are fundamental differences between the wired and wireless media that call for a more flexible design methodology for energy-constrained wireless networks.

The most fundamental difference between wired and wireless networks is in the concept of a link. This difference can be separated into two factors: *the existence of a link* and *the property of a link*. In wired networks, a link exists between a pair of nodes if and only if there is a transmission cable connecting them. Furthermore, the property of a wired link, i.e., a transmission cable, is relatively static. A wired link is usually characterized as a constant bit stream with a constant bit error probability. On the other hand, whether or not a link exists between a pair of wireless nodes depends on the transmitting/receiving decisions of these two nodes and the surrounding nodes. In particular, as long as the transmit power is large enough to overcome path loss, fading, interference, and noise, for the receiver to carry out reliable decoding, data can be transmitted between the two nodes. In other words, the existence of a wireless link is not a binary variable, rather, it depends on control decisions of nodes. In terms of link property, the wireless link is much more flexible than the wired counter part. The transmission rate and bit error probability of a wireless link can be varied by varying the transmit power.

The difference in the concept of a link makes the design and control of wireless networks much more dynamic and allows for much richer layer interactions, relative to the wired networks. For example, transmission decisions at the physical and data link layers of a wireless network can change the network topology. This in turn can affect the routing operation of the network layer. In the other direction, routing and scheduling decisions at the network and data link layers determine how multiple nodes transmit and receive data. This can affect the interference level and the link quality of the physical channel. The close interactions among different layers in a wireless network need to be carefully handled and at the same time, can be taken advantage of. To do so requires a more flexible design methodology which allows stronger interaction between layers in the protocol stack.

Another fundamental difference between wired and wireless design is in the transmission coverage. Links in wired networks are essentially point-to-point while wireless transmission is point-to-multipoint. In particular, due to the broadcast property, when one wireless node transmits, multiple nodes within the coverage of its antenna can receive the data. On one hand, this may cause unwanted interference which requires careful power control to mitigate. On the other hand, the wireless broadcast property can also be exploited to improve performance and conserve energy. Let us discuss some ideas that exploit the wireless broadcast property next.

Multicasting is an important problem of the network layer. Essentially, data need to be transmitted from a source to a set of nodes in the network. The idea of exploiting the broadcast nature of wireless media for the problem of multicasting in wireless ad hoc networks has been considered in [WNE00, SSZ01, DMS+03]. In particular, by observing that when one node transmits, the data reach multiple nodes, the number of required transmissions can be reduced to conserve energy. The wireless broadcast property also offers an opportunity for node to cooperate in routing. In particular, when a source transmits data to a destination, the surrounding nodes that receive the broadcast data can assist the transmission in different ways such as amplify and forward, decode and forward, and compress and forward. In Chapter 6, we will show how the wireless broadcast advantage can be exploited at the MAC layer for nodes in a sensor networks to jointly compress their data and conserve transmission energy.

Last but not least, an important characteristic of the wireless channel which differentiates it from the wired link is the time-varying channel gain. Due to node mobility, the channel condition between a pair of wireless nodes varies over time. Different effects such as pathloss, shadowing, and multipath fading, result in changes in the channel quality. The effects of this time-varying characteristic are twofold. Firstly, it require the control scheme to be adaptive to the fluctuation in the channel. Secondly, the changes in link condition will lead to changes in network topology. These changes will inevitably affect the operation of the whole network protocol stack. We can either fight fading or exploit fading. In fact it is shown that fading introduces a form of multiuser diversity, that can be exploited by allocating the bandwidth to the user with good instantaneous channel condition [KH95, TH98a, TH98b].

All the above characteristics, coupled with the need to conserve energy for wireless nodes, make it important to allow more interdependencies, more information sharing, and more flexibility in the design of energy-constrained wireless networks. This motivates the concept of *cross-layer design*. In general, cross-layer design is used to refer to the design approach in which protocols at different layers of the network architecture are designed in an integrated manner, with their dynamics and interdependencies being taken into account. For a detailed and concrete definition of cross-layer design, please refer to [Vin05]. In summary, the author of [Vin05] classifies cross-layer design into one of the three categories, i.e., cross-layer design based on information sharing across layers, cross-layer design based on vertical optimization of multiple protocols, and finally cross-layer design based on combining two or more adjacent layers.

Before moving on, it is important to note that cross-layer design is not only motivated by the characteristics of the wireless media. Other factors such as stochastic data arrivals, limited memory and bandwidth, and the need to guarantee quality of service (QoS) also play important roles. In fact, it is the combination of all the variations and constrains at multiple layers of wireless networks that gives rise to cross-layer design.

1.3 Thesis Focus and Contributions

This thesis focuses on cross-layer design for the first two layers of the network protocol stack, i.e., the physical (PHY) layer and the data link layer. In particular, we study different cross-layer scheduling/transmission strategies that achieve good performance, in terms of the system throughput or lifetime, while conserving energy. As a note, within the data-link layer, we mainly deal with the operation of the medium access control (MAC) sublayer. Therefore, in this thesis, we use the term "MAC layer" to refer to the MAC sublayer in the OSI model.

We note that cross-layer design for the MAC and PHY layers are an important topic due to the following reasons. First of all, in wireless networks, a large portion of energy consumption is due to data transmitting/receiving activities, which are directly controlled by scheduling/transmission schemes at the MAC and PHY layers. Secondly, as has been discussed, the variations of different parameters of the MAC and PHY layers, such as data traffic, buffer occupancies, and channel conditions, and the different concept of a wireless link are the major motivations for cross-layer design.

Our work can be divided into three main problems. We start with the first problem, which focuses on cross-layer adaptive transmission in a single-user scenario. Then in the second problem, we consider cross-layer joint adaptive scheduling/transmission in a multiple access scenario. The first and second problems are relevant in a wide range of energy-constrained networks, including cellular networks, WLANs, and wireless ad hoc networks. Finally, in the third problem, we consider a problem of combining scheduling, broadcasting, and data compression specifically for spatially correlated sensor networks. The three problems are discussed next.

1.3.1 Problem 1: Cross-layer Adaptive Transmission for Single-user Systems

We consider a discrete-time single-user system with stochastic data arrival and time-varying channel condition. Time is divided into slots of equal length and during each time slot, data packets arrive to a finite-length buffer according to some stochastic distribution. When the buffer is full, all arriving packets are dropped and considered lost. Packets are transmitted out of the buffer to a receiver over a time-varying wireless channel. The channel is represented by a finite state Markov channel (FSMC). Assume that, together with the statistics of the data arrival process and the channel variation, instantaneous buffer occupancy and channel condition are known to the transmitter and receiver. Our objective is to vary the transmit power and rate according to the buffer and channel conditions so that the system throughput is maximized, subject to an average transmit power constraint. Here the system throughput is defined as the rate of successful packet transmission. In other words, the system throughput is equal to the rate of packet arrival subtracting the rate of packet loss due to buffer overflow and transmission errors. We also consider the case when the transmit power and rate can only be chosen based on some partial observation of the buffer occupancy and channel state.

Conventional link adaptation problem only adapts the transmission parameters, i.e, power and rate, according to the condition of the time-varying channel. On the other hand, apart from the channel condition, our adaptive transmission schemes take the data arrival statistics and buffer occupancy into account.

This implies that the transmission parameters, which are the parameters of the PHY layer, are adapted to some parameters of the MAC layer. Therefore, the resultant adaptive transmission schemes can be classified as cross-layer.

In the context of link adaptation, this problem is directly related to works concerning capacity of time-varying channel with channel side information at the transmitter and receiver [GV97, GC97, ZW02]. In the context of cross-layer adaptive transmission, our work is closely related to the works in [CC99, SRB01, BG02, HGG02, GKS03, RSA04]. We defer the discussion of the related works until Chapters 2, 3 and 4.

The novelty and contributions of the work done for this problem can be summarized as follows.

- We formulate the problem of buffer and channel adaptive transmission for maximizing the system throughput, subject to an average transmit power constraint. In particular, our throughput definition incorporates effects of data arrival, buffer overflow, and transmission errors.

- We consider the throughput maximization problem under two different scenarios, i.e., when transmission is subject to a fixed bit error rate (BER) constraint and when the BER constraint is relaxed. In both scenarios, we show how optimal buffer and channel adaptive transmission policies can be obtained using dynamic programming.

- We identify an interesting and important structural property of the throughput maximizing policies, i.e., for certain correlated channel model, the optimal transmit power and rate can increase as the channel gain decreases toward outage. This is in sharp contrast to the well known water-filling structure of the transmission policy that achieves information theoretic capacity of a time-varying channel.

- We identify different practical scenarios under with the transmit power and rate can only be adapted to partial observations of the buffer and channel conditions. In those cases, we show how buffer and channel adaptive transmission can still be carried out.

The above results are discussed in Chapters 3 and 4. In particular:

- Chapter 3 is for the case when a complete observation of the instantaneous channel and buffer state information is available.

- Chapter 4 is for the case when only a partial observation of the system state is available.

1.3.2 Problem 2: Cross-layer Adaptive Scheduling / Transmission in Multiple-access Systems

In this problem, we consider a discrete-time, multiple-access scenario in which a group of nodes (users) share a common wireless channel to transmit data packets to a center node. This can be regarded as the extension of the first problem to the multiple-access scenario. Again, during each time slot, data packets arrive to the finite-length buffers of transmitting nodes according to some stochastic distribution. All buffers are finite in length and packets arriving to a full buffer are lost. For each time slot, two control decisions need to be made, i.e., a scheduling decision which assigns the common channel to one of the nodes and a transmission decision which sets the transmit power and rate for the scheduled node. All scheduling/transmission policies employed must satisfy the average transmit power constraint of each node. The objective is to adapt the scheduling and transmission decision according to the buffer and channel conditions so that the total system throughput is maximized, subject to each user average transmit constraint.

It is clear that this problem belongs to cross-layer design as i) the scheduling and transmission schemes are designed in an integrated manner and ii) the parameters from both layers, i.e., the data arrival statistics, buffer occupancies, channel statistics, and channel gain are all taken into account when making scheduling and transmission decisions.

In the context of maximizing the total system throughput, this problem is related to the work in [KH95], which concerns the sum-of-rate capacity of a multiple-access system, with channel side information at the transmitters and receiver. We will review the result of [KH95] in Chapter 2, Section 2.2.2. In the context of adapting the scheduling/transmission decisions to both buffer and channel conditions, our work is related to [TE93, AKR+01, SS02b, NMR03, LBH03, AKR+04]. These related works will be discussed in Chapter 5.

The contributions of this work are as follows.

- We formulate an optimization problem to find optimal cross-layer adaptive scheduling/transmission policies that maximize the system throughput of a multiple access system, subject to some average power constraints for all users.
- We show how MDPs can be formulated to obtain optimal as well as sub-optimal adaptive scheduling/transmission policies.

- By analyzing the performance and complexity of different class of adaptive scheduling/transmission policies, we come up with a design guideline, that can be used to determine the appropriate adaptive policy given a particular system setting.

The above results will be discussed in detail in Chapter 5.

1.3.3 Problem 3: Combining Scheduling, Broadcasting, and Data Compression in Sensor Networks

We note that the first and second problems described above focus heavily on adapting to different sources of variations in the parameters of the MAC and PHY layers. The problems considered in these two problems are also relevant to a wide range of energy-constrained networks, from cellular systems to WLANs to wireless ad hoc networks. The third problem we consider is specific to the scenario of spatially correlated wireless sensor networks. Through this work, we demonstrate that cross-layer design is still highly beneficial at the MAC and PHY layers, even when there are no variation and randomness in the system parameters.

We consider a cluster-based wireless sensor network in which sensors are organized into clusters, each cluster is responsible for monitoring a geographical area. The sensing activity is periodic, i.e., time is divided into data-gathering round and during each round, each sensor collects a fixed amount of data from the monitored field. The collected data must be transmitted directly from sensors to the corresponding cluster head. Here we assume that, within each cluster, the distance between sensors and the cluster head is short and signal strength is only affected by the free-space path loss. This means that for each sensor, both the data arrival process and channel condition are static.

Suppose that during each data gathering-round, the data collected by different sensors within the same cluster are correlated. We propose a novel approach that exploits the broadcast nature of the wireless medium so that, when one node transmits its collected data, other nodes in the same cluster can receive and use the data in compressing their own data. By doing so, they reduce the amount of data transmitted to the cluster head and conserve energy. Based on this approach, we formulate an optimization problem in which the scheduling, broadcasting, and compression decisions are made in order for sensors to collaborate in joint source compressing and conserve energy.

This problem is closely related to the works concerning joint source compression, especially distributed source coding [CPR03, ANJ05]. The idea of combining scheduling and data compression is also similar to the idea of combining routing and data compression, proposed in [SS02a]. In a broader context, this problem is based on the idea of exploiting the broadcast nature of the wireless media. Earlier works in this area include [WNE00, SSZ01, DMS+03]. These related works will be discussed in details in Chapter 6.

The novelty and contributions of this problem can be summarized as follows.

- For spatially correlated sensor networks, we propose a novel approach called collaborative broadcasting and compression (CBC), i.e., when one sensor transmits its collected data to a central node, surrounding sensors can catch the transmitted data and use them to compress their own data and therefore conserve transmission energy.

- We show how to solve for an optimal collaborative scheduling / broadcasting / compression scheme that follows the CBC approach to maximize the lifetimes of nodes in a cluster-based sensor networks.

- Finally, a heuristic algorithm, which performed well and can be obtained at lower complexity, was also proposed.

This problem will be discussed in detail in Chapter 6.

1.4 Organization of Thesis

In Chapter 2, we discuss our general system model and introduce different cross-layer scheduling and transmission strategies that will be studied in the rest of the thesis. In Section 2.1, we define the models of data arrival processes, the finite state Markov channels. Important results concerning the information capacity of time-varying channel, with channel side information available at the transmitter and receiver, are reviewed in Section 2.2. These results will be referred to in Chapters 3, 4, and 5. In Section 2.3, we discuss the need to take into account not only the channel conditions but also the buffer occupancies and data arrival statistics. This motivates our buffer and channel adaptive scheduling and transmission problems. Finally, in Section 2.4, we discuss a cross-layer scheduling, transmission, and data compression approach that can be applied to a sensor system with deterministic data arrivals and channel conditions. This approach will be studied in details in Chapter 6.

In Chapter 3, we study the problem of cross-layer adaptive transmission for single-user systems. The important assumption made in Chapter 3 is that the transmitter and receiver have a perfect knowledge of the instantaneous buffer occupancy and channel state for making transmission decisions. We start by reviewing related works in Section 3.1. Then, a concrete definition of the buffer and channel adaptive transmission problem is given in Section 3.2. We consider the problem under different scenarios, when a BER is always required (Section 3.3) and when this constraint is relaxed (Section 3.4). In Section 3.3.2, we present important result concerning the structural property of the optimal buffer and channel adaptive transmission policies. In Section 3.5, numerical results are also obtained to illustrate the performance of our cross-layer adaptive transmission policies.

In Chapter 4, we continue studying the single-user problem for scenarios when the control decisions can only be made based on some partial observation of the buffer occupancy and channel state. As discussed in Section 4.1, partial observation of the system state includes delayed and/or imperfectly estimated channel gain and quantized buffer occupancy. In Section 4.2, general approaches for buffer and channel adaptive transmission under imperfect SSI are discussed. In Section 4.3, we show that optimal adaptive policies can be obtained when some delayed but error-free channel state information is available. When this is not possible, we discuss various heuristics that achieve good performance (Section 4.4). Numerical results are provided in Section 4.5 to support our theoretical development. We note that the reader can skip this chapter and move on with Chapter 5 without loss of continuity.

In Chapter 5, the problem of cross-layer adaptive scheduling/transmission in a multiple-access scenario is studied. In Section 5.1, we discuss related works. The problem of cross-layer adaptive scheduling/transmission for maximizing the system throughput is described in Section 5.2. In Section 5.3, we show how an optimal joint adaptive scheduling/transmission policy can be obtained. In Section 5.4, we briefly discuss a class of statistic oblivious scheduling policies. These class of policies do not take the statistics of the data arrival and channel into account. In Sections 5.5 and 5.6, max-gain scheduling optimal transmission policies and round-robin scheduling optimal transmission policies are respectively considered. The performance of these two classes of suboptimal policies are studied numerically in Section 5.7. Hybrid scheduling optimal transmission is discussed in Section 5.8.

Chapter 6 is for the problem of combining scheduling, broadcasting, and data compression in spatially correlated sensor networks. We note that for the sake of understanding, the reader can go straight to this chapter while skipping Chapters 3, 4, and 5. In Section 6.1, we motivate the idea of exploiting the wireless broadcast property for sensors node to share data and carry out joint source compression. Section 6.2 is where related works are discussed. In Sections 6.3 and 6.4, the system models and general approach are introduced. The problem of combining scheduling, transmission, and joint source compression for maximizing sensors' lifetimes is defined and solved in Sections 6.5 and 6.6 respectively. In Section 6.7, a heuristic scheme which can be obtained at low complexity and achieves near optimal performance is presented. Some reflections on the design approach is given in Section 6.8. In Section 6.9, numerical results are presented to support our analysis.

Finally, in Chapter 7, we conclude this thesis by summarizing our main results, drawing important conclusions, and outlining possible avenues for future research.

CHAPTER 2

CROSS-LAYER SCHEDULING AND TRANSMISSION STRATEGIES

In this chapter, we start by discussing important system components that influence the design of scheduling and transmission strategies for energy-constrained wireless networks. These factors include the stochastic data arrival processes, finite-length buffers, and time-varying channels. Next, adaptive scheduling and transmission policies which achieve the information theoretic capacity of time-varying channels are reviewed. These adaptive policies only take into account the channel conditions while ignoring the dynamics of the data arrival processes and buffer occupancies. We then motivate the need to adapt not only to the channel conditions but also to the buffer occupancies and introduce our cross-layer adaptive scheduling and transmission problems. These problems will be studied in detail in Chapters 3, 4, and 5. Finally, we discuss a cross-layer scheduling, transmission, and data compression approach that can be applied to a sensor system with deterministic data arrivals and channel conditions. This approach will be studied in detail in Chapter 6.

2.1 General System Model

The general system model considered in this thesis can be depicted in Fig. 2.1. There are N nodes (users) that communicate with a center node over the wireless medium. N users are numbered: 1, 2, $\ldots$ N. We consider a discrete-time system in which time is divided into slots, each of length equal to T_s seconds, $T_s > 0$. Time slot i, $i \in \mathbb{N}$, denotes the time period $[iT_s,\ (i+1)T_s)$. During each time slot, data, in terms of fixed-sized packets, arrive to the buffer

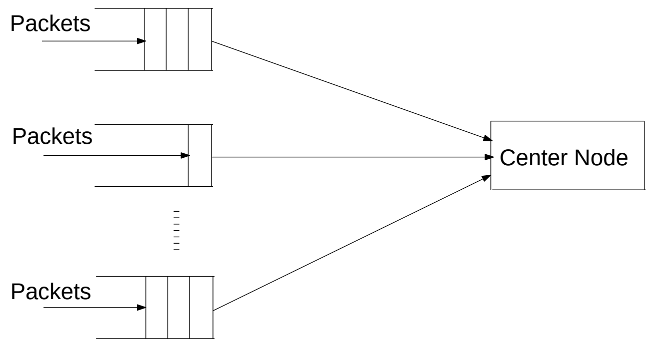

Figure 2.1: A wireless system in which multiple wireless nodes communicate their data toward a center node.

of each user. The users then transmit data from their buffers, in a first-in-first-out (FIFO) manner, over the wireless channel toward the center node.

The system model in Fig. 2.1 can fit into different scenarios of energy-constrained wireless networks discussed in Chapter 1. In a cellular system, this model represents the uplink communication from mobile terminals to a base station. In a WLAN network, Fig. 2.1 models multiple wireless nodes transmitting data toward an access point. Fig. 2.1 can also be thought as a portion extracted from a wireless ad hoc network. Finally, for sensor network applications, Fig. 2.1 depicts the scenario in which a number of sensors forward collected data toward a data aggregation/fusion center.

It is clear that when N is set to 1, we have a single-user system. The problems considered in Chapters 3 and 4 will be for the single-user system while Chapter 5 will deal with the multiple-access scenario.

2.1.1 Data Arrival Processes and Buffer Dynamics

Let A_i^n denote the number of data packets that arrive to the buffer of user n during time slot i, $i \in \mathbb{N}$ and $n \in \{1,\ 2,\ \ldots\ N\}$. Depending on the specific application, the arrival process $\{A_i^n\}$ can be either deterministic or stochastic. For example, in data-logging sensor networks, each sensor collects a fixed amount of data periodically [KDN03]. On the other hand, data and multimedia traffics are usually stochastic in nature and can be modeled by different stochastic processes such as Poisson processes and Markov modulated Poisson processes ([AN98]). In Chapters 3, 4, 5, we assume that the data arrival processes are independent and identically distributed (i.i.d.) over time and across all users. However, this assumption can be relaxed and the results in those chapters can be easily extended to the case of Markov arrival processes.

Let B^n denote the size (in packets) of the buffer of user n. Also, let B_i^n denote the buffer occupancy, i.e., the number of queueing packets, of user n *at the beginning of time slot* i. We assume that packets that arrive to the buffer during time slot i are only added to the buffer *at the end of time slot* i. If there is no space left in the buffer, arriving packets are dropped and considered lost. Suppose U_i^n, $U_i^n \leq B_i^n$, is the number of packets that are emptied from the buffer of user n during time slot i, we can write

$$B_{i+1}^n = \min\{B^n,\ B_i^n - U_i^n + A_i^n\}. \qquad (2.1)$$

2.1.2 Finite-state Markov Channels

In Chapters 3, 4, 5, we consider discrete-time block-fading channels with additive white Gaussian noise (AWGN). W (in Hz) and $N_o/2$ (in Watts/Hz) denote the channel bandwidth and the noise power density respectively. The fading process seen by each user is represented by a stationary and ergodic K-state Markov chain. It is assumed that the channel stays in the same state for an entire time slot. Let G_i^n denote the channel state of user n during time slot i, we assume that the steady-state distribution as well as the transitioning probabilities of G_i^n are known at both the transmitters and the receiver. This assumption is reasonable in systems where the fading process is slow enough so that necessary statistics can be estimated at the receiver and fed back to the transmitters.

In general, finite-state Markov channels (FSMCs) are suitable for modeling slowly varying frequency-flat fading [Gud91, WM95, HGG02, BG02]. An FSMC can be constructed for a particular fading distribution by first partitioning the range of the fading gain into a finite number of sections. Then each section of the fading gain corresponds to a state in the Markov chain. Given the statistics of the fading process, the stationary distribution as well as the channel state transitioning probabilities can be determined.

As an example, let us demonstrate how a slowly-varying Rayleigh fading channel can be represented by an FSMC. The following derivations are extracted from [WM95].

First, let γ denote the instantaneous channel power gain, which is proportional to the squared envelope of the received signal. For Rayleigh fading, the probability density function (p.d.f.) of γ can be written as

$$p_\Gamma(\gamma) = \begin{cases} \frac{1}{\overline{\gamma}} \exp\left(-\frac{\gamma}{\overline{\gamma}}\right) \\ 0, \ \gamma < 0 \end{cases} \tag{2.2}$$

where $\overline{\gamma}$ is the average channel gain [Pro01].

Let the channel gain γ be partitioned into K intervals using $K+1$ levels: $0 = \gamma_0 < \gamma_1 < \ldots < \gamma_{K-1} < \gamma_K = \infty$. We say that the channel is in state $k,\ k \in \{0, 1, \ldots K-1\}$ if $\gamma_k \leq \gamma < \gamma_{k+1}$. From (2.2), the steady-state probability of channel state k is

$$p_G(k) = \exp\left(-\frac{\gamma_k}{\overline{\gamma}}\right) - \exp\left(-\frac{\gamma_{k+1}}{\overline{\gamma}}\right). \tag{2.3}$$

We still need to calculate the state transitioning probabilities for the K-state channel model. From [Jak74], the channel cross over rate, i.e., the expected number of times per second γ crosses a particular level γ_k (in either up or down direction), is

$$N_k = \sqrt{\frac{2\pi\gamma_k}{\overline{\gamma}}} f_D \exp\left(-\frac{\gamma_k}{\overline{\gamma}}\right),\ \ k \in \{0, 1, \ldots K-1\}, \tag{2.4}$$

where f_D is the maximum Doppler frequency of the channel. Given that the mobile terminal moves at speed v (meters/second) and the carrier frequency is f,

$$f_D = f\frac{v}{c}, \tag{2.5}$$

where c (in meters/second) is the speed of light. Now, if we assume that the fading is slow enough so that the channel gain stays in the same state during each time slot and state transitions after each time slot only happen between adjacent states, the state transitioning probabilities of the K-state channel after each time slot can be written as

$$\begin{aligned}
&P_G(k,k-1) = \frac{N_k T_s}{p_G(k)},\ k \in \{1,2,\dots K-1\},\\
&P_G(k,k+1) = \frac{N_{k+1} T_s}{p_G(k)},\ k \in \{0,1,\dots K-2\},\\
&P_G(k,k) = 1 - P_G(k,k-1) - P_G(k,k+1),\ k \in \{1,2,\dots K-2\},\\
&P_G(0,0) = 1 - P_G(0,1),\\
&P_G(K-1,K-1) = 1 - P_G(K-1,K-2).
\end{aligned} \tag{2.6}$$

Now the K-state Markov channel model has been completely specified with the steady-state probability in (2.3) and the state transition probabilities in (2.6).

2.2 Capacity-achieving Strategies for Fading Channels

In this section, we review results concerning the information theoretic capacity of time-varying wireless channels, with channel side information available at both the transmitter and receiver. We consider two scenarios, i.e., for single-user systems and for multiple-access systems. Note that as the information theoretic capacity is of concern, the data arrival statistics and buffer condition are not taken into consideration.

2.2.1 Single-user Scenario

Consider a single-user system in which a transmitter sends data to a receiver over a time-varying wireless channel. As before, the instantaneous channel power gain is denoted by γ. The probability distribution of γ is $p_\Gamma(\gamma)$.

Assuming that the instantaneous value of γ is available at both the transmitter and the receiver, a power control policy is defined as a function $P(\gamma)$ which sets the transmit power when the channel gain is γ. Suppose that all power control policies employed must satisfy the average transmit power constraint

$$\int_\gamma P(\gamma)p_\Gamma(\gamma)d\gamma \;\le\; \overline{P}. \tag{2.7}$$

We are interested in the capacity of this system, i.e., the maximum transmission rate that can be achieved with some power control and coding schemes such that the probability of error is arbitrarily small. In [GV97], Goldsmith and Varaiya give the following definition for the fading channel capacity and subsequently prove a channel coding theorem and converse.

Definition 2.2.1. *([GV97]) Given the average power constraint* (2.7), *define the time-varying channel capacity by*

$$C(\overline{P}) = \max_{P(\gamma):\ \int_\gamma P(\gamma)p_\Gamma(\gamma)d\gamma \;\le\; \overline{P}} \int_\gamma W\log_2\left(1+\frac{P(\gamma)\gamma}{N_oW}\right)p_\Gamma(\gamma)d\gamma. \tag{2.8}$$

In particular, it is shown in [GV97] that the power control policy which maximizes (2.8) exhibits the following interesting structure.

$$\frac{P(\gamma)}{N_oW} = \begin{cases} \frac{1}{\gamma^*} - \frac{1}{\gamma},\ \gamma \ge \gamma^* \\ 0,\ \gamma < \gamma^*. \end{cases} \tag{2.9}$$

Equation (2.9) tells us that there is a cutoff value γ^* below which no transmission should be carried out. Above this cutoff value, the power allocation follows a *water-filling* ([Gal68, BV04]) structure in time, with more transmit power (and rate) being allocated when the channel gain increases. The value of γ^* depends on the channel gain distribution and the power constraint through

$$\int_{\gamma^*}^{\infty} N_oW\left(\frac{1}{\gamma^*}-\frac{1}{\gamma}\right)p_\Gamma(\gamma)d\gamma \;=\; \overline{P}. \tag{2.10}$$

Substituting (2.9) into (2.8) gives us the capacity of fading channel, with channel side information at both the transmitter and the receiver. The coding/decoding scheme which achieves this capacity is described in [GV97]. The main idea is to multiplex multiple coding and modulation schemes, each optimized for a particular fade level. The resultant coding/decoding scheme is both variable-power and variable-rate.

We note that the results discussed above are for the information theoretic capacity, which can not be achieved with practical coding/decoding schemes. In [GC97], Goldsmith and Chua consider a similar problem of communicating over a time-varying channel, but in a practical setup. Adaptive transmission is based on a variable-power variable-rate M-ary quadrature modulation (MQAM) scheme. In particular, by fixing the symbol rate while varying the signal constellation size, different transmission rates can be achieved. Similar to the information theoretic setup, the transmission rate and power are varied based on instantaneous channel gain. The objective is to find an adaptive MQAM scheme that maximizes the average transmission rate, subject to the constraints on average transmit power and bit error rate. It is interesting to see that the optimal adaptive MQAM scheme that maximizes the expect transmission rate also follows the water-filling structure [GC97]. In particular, more transmit power (and rate) is allocated when the channel gain increases.

2.2.2 Multiple-access Scenario

The capacity of a multiple-access system is characterized by its capacity region, i.e., the set of all possible rate vectors that can be supported by the system with arbitrarily small probability of error. Within this capacity region, an important performance metric is the sum-of-rate capacity, i.e., the maximum total achievable rates for all users.

Let $\underline{\gamma}$ be the vector of instantaneous channel gain of N users, with γ^n being the channel gain for user n, $n \in \{1, 2, \dots N\}$. Again, we assume that the instantaneous value of $\underline{\gamma}$ is available at the transmitters and receiver. Let $P^n(\underline{\gamma})$ be a power control scheme that set the transmit power for user n when the instantaneous channel gain is $\underline{\gamma}$. In [KH95], Knopp and Humblet study the problem of maximizing the sum-of-rate capacity of a multiple-access system, subject to the average transmit power constraint of each node. In particular, the objective is to maximize

$$C(\overline{P}) = \int_{\gamma^1}\int_{\gamma^2}\cdots\int_{\gamma^N} W\log_2\left(1+\sum_{n=1}^{n}\frac{P^n(\underline{\gamma})\gamma^n}{N_oW}\right)p(\underline{\gamma})d\underline{\gamma}, \tag{2.11}$$

subject to

$$\int_{\gamma^1}\int_{\gamma^2}\cdots\int_{\gamma^N} P^n(\underline{\gamma})p(\underline{\gamma})d\underline{\gamma} \le \overline{P}^n, \ \forall n \in \{1, 2, \dots N\}. \tag{2.12}$$

Note that in (2.11) and (2.12), $\overline{P} = (\overline{P}^1, \overline{P}^2, \dots \overline{P}^N)$ where $\overline{P}^n$ is the average power constraint of user n.

It is shown in [KH95] that the power allocation scheme that maximizes (2.11) has the following form.

$$\frac{P^n(\underline{\gamma})}{N_oW} = \begin{cases} \frac{1}{\gamma^{n*}} - \frac{1}{\gamma^n}, \ \gamma^n \ge \gamma^{n*}, \ \frac{\gamma^n}{\gamma^{n*}} > \frac{\gamma^m}{\gamma^{m*}}, \ m \ne n \\ 0, \ \text{otherwise.} \end{cases} \tag{2.13}$$

From 2.13, it can be seen that at each time instance, the channel is allocated to at most one user. The user who is assigned the channel must have the relatively best channel gain. Moreover, for the selected user, transmit power is again allocated according to a water-filling structure in time.

2.3 Taking Arrival Statistics and Buffer Occupancies into Account

The results regarding the information theoretic capacity of single-user systems and the sum-of-rate capacity of multiple-access systems highlight the importance of adapting to the instantaneous channel gain. In particular, these results motivate the intuitive approach of exploiting favorable channel conditions when communicating over time-varying wireless channels. Here, channel variations happen over time and across users. By definition, the power and rate adaptive policies described in Section 2.2 achieve the corresponding capacities. However, as we will illustrate next, when other factors of a practical system are taken into account, these capacity-achieving policies may not guarantee the best system throughput.

Consider a single-user system with stochastic data arrival, a finite-length buffer, and a time-varying channel (as a special case of the general model described in Section 2.1). If the capacity-achieving adaptive power and rate is employed, we will transmit at higher power and rate when the channel gain increases. However, due to the random data arrival process and limited buffer space, there can be time when the channel is good but the buffer is near empty and prohibits transmission at a high rate. There can also be time when the channel condition is not favorable but the buffer is close to overflow and requires transmission at high rate. This suggests the importance of taking into account not only the channel condition, but also the statistics of the data arrival process and the buffer occupancy when making transmission decisions.

Similar situations arise in multiple-access systems with stochastic data arrival and limited buffers. According to the policy derived by Knopp and Humblet, the common channel is always assigned to the users with the relatively best channel gain. However, due to random arrival and limited buffers, user with the best channel condition may have a near empty buffer. In that case, it is wiser to assign the common channel to a user with a less favorable channel condition but near-overflow buffer.

Finally, we note that the results regarding the information theoretic capacity assume that very long codewords can be used in data transmission. Using long codewords incurs long delay at both transmitters and receivers and that can violate some delay requirements of the application. Long queueing delay at the transmitter buffer also leads to higher probability of buffer overflow, which directly affects the system throughput. Last but not least, with small buffers, it is also not possible to use codewords that are long enough to guarantee arbitrary small error probability. As a result, all transmission will suffer some positive error probability.

2.3.1 System Throughput

The above discussion motivates us to study scheduling/transmission strategies under some performance metric that is more meaningful to the applications. Depending on different applications and system scenarios, there can be different suitable performance metrics. These metrics include, but not limited to, average queueing delay, deadline violation rate, buffer overflow probability, packet error probability, and throughput. However, we are more interested in the system throughput, as this metric allows us to relate to the results concerning the information theoretic capacity.

For our system, a definition of the system throughput should take into account the effects of stochastic data arrival processes, finite buffer lengths, and transmission errors. We observe that stochastic data arrival and finite buffer lengths lead to packet loss due to buffer overflow while transmission errors can result in erroneous packets being discarded. Therefore, we propose the following definition for the system throughput:

$$throughput = arrival\_rate - overflow\_rate - error\_rate. \qquad (2.14)$$

Here *arrival_rate* is the long term average rate at which data arrive to the buffers, *overflow_rate* is the rate at which packets are dropped due to buffer overflow, and *error_rate* is the rate at which packets are discarded due to transmission errors. Note that in multiple user systems, the rates are summed up across all users.

2.3.2 Buffer and Channel Adaptive Policies

For a single-user system, let $\boldsymbol{S}_i = (B_i, G_i)$ denote the system state at time i, $i = 0, 1, \ldots$. Here B_i is the number of packets queueing in the buffer at the beginning of time slot i while G_i is the channel state during time slot i. We are interested in adaptive policies that adapt the transmit power and rate according to the system state $\boldsymbol{S}_i$. Let P_i and U_i be the transmit power and rate for time slot i. We will study the following throughput maximization problem.

Throughput Maximization Problem (for single-user systems): *For each time slot i, based on the system state $\boldsymbol{S}_i$, select the transmit power P_i and rate U_i so that the system throughput is maximized, subject to the average transmit power constraint.*

The above throughput maximization problem will be studied in Chapters 3 and 4, under different scenarios. As it will be shown, the optimal buffer and channel adaptive transmission policies that maximize the system throughput may exhibit a structural property that is very different from that of capacity-achieving policies described in Section 2.2.1.

For the multiple-access system in Fig. 2.1, the system state includes the buffer and channel states for all N users, i.e.,

$$\boldsymbol{S}_i = (B_i^1, B_i^2, \dots B_i^N,\ G_i^1, G_i^2, \dots G_i^N).$$

There are two decisions to make in each time slot, i.e., i.e., a scheduling decision which assigns the common channel to one of the nodes and a transmission decision which sets the transmit power and rate for the scheduled node. In Chapter 5, we will study the following problem.

Throughput Maximization Problem (for multiple-access systems): *For each time slot i, based on the system state $\boldsymbol{S}_i$, select a user to access the channel and for this user, assign the transmit power P_i and rate U_i so that the system throughput is maximized, subject to average transmit power constraint for each of the N users.*

2.4 A Cross-layer Strategy under Deterministic Data Arrival and Deterministic Channel

So far, we have motivated cross-layer scheduling/transmission schemes that adapt to the randomness and time variations of the data arrival processes and fading channels. These schemes will be studied in detail in Chapters 3, 4, 5.

In this section, we will introduce another cross-layer scheme, which is applied for system with deterministic data arrival and channel conditions. This can be considered as a different type of cross-layer design, which focuses on the cooperation of protocols at different layers in the protocol stack. This scheme will later be studied in Chapter 6.

2.4.1 A Periodic Sensing Scenario with Spatial Data Correlation

To begin with, we note that Fig. 2.1 can be used to depict a sensing application scenario in which multiple sensors transmit data toward a center node who is responsible for data aggregation/fusion. Let us consider a periodic sensing scenario in which sensors collect a fixed amount of data during each time slot. At the end of time slot, all sensors need to communicate that data toward the common node.

An important characteristic in sensing application is that data collected by different sensors can be correlated. This is particularly true for sensors located close to one another. In that case, if sensors node can collaborate with each other, they can jointly compress data before transmission. That can help reduce transmission energy. We will address compression of correlated information source next.

2.4.2 Compression of Correlated Information Sources

Let us consider two information sources that generate correlated discrete random variables X and Y. Variable X takes values from a set $\mathcal{X}$ with a probability distribution $p_X(x)$. Similarly, Y takes values from a set $\mathcal{Y}$ with a probability distribution $p_Y(y)$. Furthermore, the correlation between X and Y is specified by a probability distribution $p_{XY}(x, y)$. We supposed that the two information sources are compressed by two encoders and then decoded by a common decoder.

If X and Y are encoded/decoded independently, Shannon's Theorem states that the average number of bits per source symbol required to noiselessly encode X and Y are $H(X)$ and $H(Y)$ respectively, where

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log_2 p_X(x) \quad \text{and} \quad H(Y) = -\sum_{y\in\mathcal{Y}} p_Y(y)\log_2 p_Y(y) \tag{2.15}$$

are the entropies of variables X and Y.

However, the correlation between X and Y can be exploited to reduce the total number of bits required to reliably encode them. In particular, if the encoders of X and Y can access each other's information, X and Y can be compressed without loss to the rates R_X and R_Y that satisfy:

$$R_X \geq H(X), \tag{2.16}$$

$$R_Y \geq H(Y), \tag{2.17}$$

$$R_X + R_Y \geq H(X, Y), \tag{2.18}$$

where $H(X, Y)$ is the joint entropy of X and Y and can be calculated as:

$$H(X, Y) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x, y)\log_2 p_{XY}(x, y). \tag{2.19}$$

As an example, suppose that the encoder of Y explicitly knows X. Then X and Y can be losslessly compressed at rates $R_X = H(X)$ and $R_Y = H(X, Y) - H(X) = H(Y|X)$. Note that

$$H(Y|X) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x, y)\log_2 p_{Y|X}(y|x) \tag{2.20}$$

where $p_{Y|X}(y|x)$ is the conditional probability distribution of Y given X.

In the above discussion, we have assumed that the encoders of X and Y can share information with each other. However, in [SW73], Slepian and Wolf presented an important result, which showed that X and Y can be encoded and decoded with arbitrarily small probability of error at rates R_X and R_Y satisfying (2.16), (2.17), (2.18), even when the two encoders work independent to each other. As long as the two encoders know the correlation statistics of X and Y, noiseless compression can be carried out. The encoding/decoding scheme proposed by Slepian and Wolf is usually termed *distributed source coding.*

2.4.3 Exploiting Wireless Broadcast Property for Data Compression

Getting back to the sensing scenario described in Section 2.4.1, the theories of compression of correlated sources motivate us to allow sensors that collect correlated data to carry out joint data compression. As discussed in Section 2.4.2, joint data compression can be done by either letting sensors to explicitly share their collected data, or following the distributed source coding approach of Slepian and Wolf.

In Chapter 6, we propose a novel approach that allows sensors to carry out joint data compression based on explicitly sharing their collected data. One advantage of encoding based on explicit information, over distributed source coding, is that the encoding/decoding schemes can be much simpler [SS02a]. The core idea of our approach is as follows. Since wireless transmission is inherently broadcast, when one sensor transmits its collected data, other sensors in its coverage area can receive the transmitted data. These sensors can therefore utilize the data they overhear from other nodes in compressing their own data so that transmission energy can be conserved. Based on this idea, we proposed the following approach.

Collaborative Broadcasting and Compression (CBC): *Given a set of sensors transmitting correlated data to a center node, schedule their data transmission and reception so that joint data compression based on explicit information can be carried out, with the objective of conserving sensors' energy and extending their lifetimes.*

From the system design point of view, the CBC approach is cross-layer in that it integrates the scheduling, transmission, reception, and data compression operations for the sensor nodes.

2.5 Summary

In summary, we presented the general system model that will be used in the studies of Chapters 3, 4, and 5. Important components of the system model, i.e., the data arrival processes, the buffer dynamics, and the time-varying channels, have been discussed.

We also reviewed policies that achieve the information theoretic capacity of single-user systems and sum-of-rate capacity of multiple-access systems, both with time-varying channels. Note that the channel side information is assumed to be perfectly available at the transmitters and receiver. The capacity-achieving policies tend to favor good channel conditions (either over time or across users) by setting higher transmit power and rate when the channel gain increases.

However, as these capacity-achieving policies are only concerned with the "maximum achievable rate", they do not take into account the data arrival statistics and buffer occupancies. This leads to sub-optimality, in terms of the actual throughput that can be achieved. This motivated us to define cross-layer adaptive scheduling/transmission problems which take into account not only the channel statistics, but also the data arrival statistics and buffer occupancies. These problems will be studied in detail in Chapters 3, 4, and 5.

As can be noted, the cross-layer adaptive scheduling/transmission problems introduced in this chapter are largely motivated by the need to adapt to the stochastic time-variations in different parameters of the MAC and PHY layers. In Chapter 6, with a system model similar to that of Fig. 2.1, we will present a cross-layer approach which is beneficial to the system even when there is no time-varying factor in the system components. To get a quick look at this problem, the reader can skip Chapters 3, 4, 5 altogether and go straight to Chapter 6.

CHAPTER 3

BUFFER AND CHANNEL ADAPTIVE TRANSMISSION: FULLY OBSERVABLE SYSTEM STATES

In this chapter, we study a problem of cross-layer adaptive transmission in a single-user scenario. This problem has been introduced in Chapter 2, Section 2.3. In our system, time is divided into slots of equal length and during each time slot, data packets arrive to a finite-length buffer according to some stochastic distribution. When the buffer is full, all arriving packets are dropped and considered lost. Packets are transmitted out of the buffer to a receiver over a time-varying wireless channel. Assume that, together with the statistics of the data arrival process and the channel variation, instantaneous buffer occupancy and channel gain are known to the transmitter and receiver. The objective is to vary the transmit power and rate according to the buffer and channel conditions so that the system throughput is maximized, subject to an average transmit power constraint. The system throughput is defined as the rate of successful packet transmission and can be calculated by subtracting the rate of packet loss due to buffer overflow and transmission errors from the packet arrival rate.

We note that, apart from the channel condition, our adaptive transmission problem takes the data arrival statistics and buffer occupancy into account. In other words, the transmission parameters of the PHY layer are varied based on some parameters of the MAC layer. Therefore, the resultant adaptive transmission schemes can be classified as cross-layer.

We first consider the case when transmission is subject to a fixed bit error rate (BER) constraint. In that case, the packet error rate (PER) is also fixed and maximizing the system throughput is equivalent to minimizing the rate of packet loss due to buffer overflow. This may be appropriate in situations where a certain quality of service is mandated by communication standards or specific user applications. When the BER constraint is relaxed, we have a trade-off between packet loss due to transmission errors and packet loss due to buffer overflow. In that case, we solve for a buffer and channel adaptive transmission policy that minimizes the total packet loss due to buffer overflow and transmission errors.

The main results of this chapter can be summarized as follows.

- We formulate a problem of adapting the transmit power and rate according to the buffer occupancy and channel gain so that the system throughput is maximized, subject to an average transmit power constraint.

- We solve the above throughput maximization problem in two scenarios, when adaptive transmission is carried out with and without a BER constraint.

- We show that for some throughput-maximizing policies, the optimal transmit power and rate can increase as the channel gain decreases toward the outage threshold. This effect is in contrast to the water-filling property of the policies that achieve the information theoretic capacity of time-varying channels.

- We present numerical results to support the theoretical development.

We note that some of the above results have been presented in [HM02, HM03, HM04a, HM05d].

The rest of this chapter is organized as follows. In Section 3.1, work related to the problem considered in this chapter is reviewed. In Section 3.2, the system model and the throughput maximization problem is defined. In Section 3.3, we study the throughput maximization problem when transmission is subject to a BER constraint. In particular, in Section 3.3.1, we describe our approach of formulating the problem as a Markov decision process (MDP) and using dynamic programming to solve it. Section 3.3.2 discusses an interesting structural property of the optimal policies. In Section 3.4, we remove the BER constraint and consider the problem of minimizing total packet loss rate due to both buffer overflow and transmission errors. In Section 3.5, we present numerical results and discussion. The chapter ends with some concluding remarks in Section 3.6.

3.1 Related Work

In the context of link adaptation, our work is related to the works by Goldsmith in [GV97, GC97]. In [GV97], the information theoretic capacity of a time-varying channel is characterized for the case in which the channel state information is available at both the transmitter and receiver. As has been discussed in Chapter 2, Section 2.2, the capacity-achieving transmission policy in [GV97] adapts the transmit power and rate according to the instantaneous channel condition. More specifically, the transmit power is allocated according to a water-filling structure in time. In [GC97], a variable-rate, variable-power M-ary quadrature amplitude modulation (MQAM) is proposed to maximize the throughput of transmission over a time-varying channel. Again, the policy that maximizes the system throughput allocates power according to a water-filling rule.

We note that the objective of our work and that of [GV97, GC97] are similar, i.e., to maximize the throughput of transmission over a time-varying channel subject to an average power constraint. However, in our study, we take into account the effects of a stochastic data arrival process, a finite-length buffer, and transmission errors and adapt the transmit power and rate not only to the channel gain but also to the buffer occupancy. With this formulation, we point out an interesting structural property of the optimal adaptive transmission policies, i.e., for certain correlated fading channel models, the optimal transmit power and rate can increase as the channel gain decreases toward outage. This is in sharp contrast to the water-filling property of the policies presented in [GV97, GC97].

In the context of cross-layer adaptive transmission, our work is closely related to the works in [CC99, SRB01, BG02, HGG02, GKS03, RSA04]. In all of these related works, similar single-user system models with stochastic data arrivals and time-varying channels are considered. In [CC99], Collins and Cruz study the problem of adapting the transmit power and rate according to the instantaneous buffer occupancy and channel state in order to minimize the average transmit power, subject to some constraints on average delay and peak transmit power. The results in [CC99] demonstrate that a good adaptive transmission policy should take into account both the channel condition and user's backlog. In [BG02] and [GKS03], the problem in [CC99] is studied under more general scenarios. In particular, the work of Berry and Gallager in [BG02] gives a thorough characterization of the tradeoff between average delay and average transmit power for the regime of asymptotically large delay. The objective of [RSA04] is similar to that of [BG02], i.e., to characterize the optimal achievable region of average transmit power and average delay. However, the system model of [RSA04] are more specific, with bounded data arrivals and only zero-outage policies are considered.

If the works in [CC99, BG02, GKS03, RSA04] deal with the average delay, the works in [SRB01, HGG02] are suitable for cases of packet transmission with strict deadlines. In [SRB01], Shabharwal et al. consider the problem of buffer and channel adaptive transmission to minimize packet loss for a system with constraints on the average transmit power and maximum queueing delay. The authors show that a small increase in the allowable delay leads to reduction in packet loss probability. In [HGG02], a similar problem is considered, with the objective of minimizing average transmit power, subject to a constraint on the probability of packet loss due to deadline expiry.

We note that our work differs from the works in [CC99, SRB01, BG02, HGG02, GKS03, RSA04] in several significant ways. First, while these related works deals with delay, the objective of our work is to maximize the system throughput which are related to buffer overflow and transmission errors. Second, in [CC99, BG02, GKS03], the authors characterize how the optimal transmission rate depends on the channel condition, however, their characterization is only for the case when the fading process is independent and identically distributed (i.i.d.) over time. In that case, the structure of the optimal policies is similar to the water-filling structure. In our work, we look at the dependency when the fading process is time correlated and show an interesting observation. In Chapter 4, we also study the problem under cases when the transmitter only has some incomplete knowledge of the system state information. Incomplete system state observation is not considered in [CC99, SRB01, BG02, GKS03, RSA04].

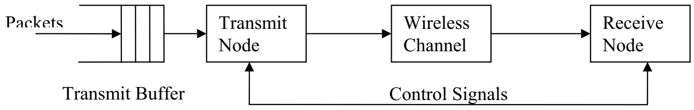

Figure 3.1: Single-user data communication system with stochastic data arrival, finite-length buffer, and time-varying channel condition. Channel and buffer conditions are signaled between the transmitter and receiver.

3.2 Problem Definition

3.2.1 System Model

The system model considered in this chapter and the following chapter is depicted in Fig. 3.1. We have a single-user data communication system with stochastic data arrival, finite-length buffer, and time-varying channel conditions. Important components of this system have been discussed in Chapter 2. The main assumptions and notations can be summarized as follows.

- Time is divided into slots of equal length of T_s seconds. Slot i, $i \in \mathbb{N}$, refers to the time period $[iT_s,\ (i+1)T_s)$.
- The number of packets arriving to the buffer during slot i is denoted by A_i. We assume that these A_i packets are only added to the buffer *at the end* of slot i. In this work, we consider the case when $\{A_i\}$ is independent and identically distributed (i.i.d.) over time so that the index i can be omitted. We note that the results can be extended to consider a time-correlated arrival process in a straightforward manner. The distribution of the number of packets arriving during each time slot is assumed known and denoted by $p_A(a)$, i.e., $p_A(a) = \Pr(A_i = a)$. The average packet arrival rate is denoted by λ (packets/second).

- All packets have the same length of L bits. The buffer can store up to B packets and if a packet arrives when the buffer is full, it is dropped and counted as a packet loss.

- We consider a discrete-time block-fading channel with additive white Gaussian noise (AWGN). We use W(Hz) and $N_o/2$(Watt/Hz) to denote the channel bandwidth and the AWGN noise power density respectively. The fading process is represented by a stationary and ergodic K-state Markov chain, with the channel states numbered from 0 to $K-1$. The channel power gain of state g, $g \in \{0, \dots K-1\}$, is denoted by γ_g. During each time slot, the channel remains in a single state. Let G_i denote the channel state during time slot i, the channel state transition probability is defined as

$$P_G(g, g') = \Pr\{G_i = g' \mid G_{i-1} = g\}. \tag{3.1}$$

  The stationary distribution of each channel state g is denoted by $p_G(g)$. We assumed that $P_G(g, g')$ and $p_G(g)$ are known for all $g, g' \in \{0 \dots K-1\}$.

- We denote the system state in slot i by $\boldsymbol{S}_i = (B_i, G_i)$, where $B_i \in \{0, \dots B\}$ is the number of packets in the buffer *at the beginning* of slot i while $G_i \in \{0, \dots K-1\}$ is the channel state *throughout* slot i.

For more details on the data arrival process, buffer dynamics, and finite-state Markov channels, please refer to Chapter 2, Section 2.1.

3.2.2 Adaptive Transmission

At the beginning of time slot i, we assume that the transmitter and the receiver have a perfect knowledge of the current system state $\boldsymbol{S}_i$. This assumption is reasonable for cases when the channel changes slowly so that its state can be reliably estimated and fed back within the control overhead. In Chapter 4, cases when the transmission decisions can only be made based on some partial system state information (SSI) will be considered.

We assume that, based on $\boldsymbol{S}_i$, the transmitter can vary its transmit power and rate during each time slot. For time slot i, let P_i(Watts) and U_i(packets/slot) denote the transmit power and rate respectively. We require that $U_i \in \{0, 1, \ldots B_i\}$ and $P_i \in \boldsymbol{\mathcal{P}}$ where $\boldsymbol{\mathcal{P}}$ is the set of all power levels that the transmitter can operate at. The most general case is when $\boldsymbol{\mathcal{P}}$ is the set of all non-negative real numbers $\mathbb{R}^+$. A pair (U_i, P_i) is called a control action for time i.

Note that there are various ways for the transmitter to change its transmission rate U_i. It can be done by changing the channel coding scheme [Vuc91], i.e. by encoding data bits in the buffer using different code rates while keeping the transmission rate for the coded bits fixed. U_i can also be varied by keeping the symbol rate fixed and changing the signal constellation size of a modulator [WS95, GC97, HM02, HM03]. As an example, in the IEEE.802.11 standards, different transmission rates are achieved by combinations of different coding and modulation schemes.

An adaptive scheme that varies transmit power and rate based on the buffer occupancy and channel state can be implemented as lookup table. At the beginning of slot i, the system updates $\boldsymbol{S}_i$, which gives an index in the lookup table of transmission parameters. Advances in communication devices, especially in software radio, will allow this adaptive functionality to be efficiently implemented in transceivers. As an example, in a software radio based system, adaptive policies, i.e., lookup tables, can be stored at a base station or downloaded over the air-interface and stored at mobile nodes. Also, by simply changing some parameters, coding and modulation schemes can be reconfigured when they are implemented in software.

3.2.3 Transmission Errors

We use $P_b(g, u, P)$ to denote the function that gives the bit error rate (BER) when the channel state is g and the transmit power and rate are P and u respectively. The packet loss rate, as a function of the BER, depends on the specific packet error-correcting scheme being implemented. In this work, we suppose that a packet is in error if at least l out of its L bits are corrupted. Then we can characterize the packet error probability in terms of u, g, P as

$$P_p(g, u, P) = \sum_{j=l}^{L} \binom{L}{j} P_b(g, u, P)^j \big(1 - P_b(g, u, P)\big)^{(L-j)}. \qquad (3.2)$$

We note that the function $P_b(g, u, P)$ depends on the specific coding, modulation, and detection schemes used. As an example, let us vary the transmission rate by varying the constellation size of an M-ary quadrature amplitude modulator (MQAM) while fixing its symbol rate. When the channel is in a particular state g, the fading gain is γ_g. With the channel state available at both transmitter and receiver, the channel can be treated as AWGN. From [FS83], assuming ideal coherent phase detection, the BER for a particular transmit power P and rate u bits per QAM symbol can be upper-bounded by

$$P_b(g, u, P) \leq 2 \exp\left(-1.5 \frac{P\gamma_g}{WN_o(2^u - 1)}\right). \qquad (3.3)$$

For $u \geq 2$ and $0 \leq SNR \leq 30$ dB, a tighter bound for the BER is given by

$$P_b(g, u, P) \leq 0.2 \exp\left(-1.5 \frac{P\gamma_g}{WN_o(2^u - 1)}\right). \qquad (3.4)$$

We will later use these upper bounds to approximate the transmit power needed to meed a BER constraint.

3.2.4 Throughput Maximization Problem

From the system point of view, an important performance metric is the rate at which data packets are successfully transmitted. In this work, we assume that all packets that are dropped due to buffer overflow and all packets that are discarded due to transmission errors are lost. However, our formulation can also be extended to account for the retransmission of erroneous packets. We give the following definition of the system throughput.

Definition 3.2.1. *The system throughput is the long term average rate at which packets are successfully transmitted. For an average packet arrival rate* λ*, a buffer overflow probability* P_{of}*, and a packet error rate* $\overline{P_p}$*, the system throughput can be calculated as:*

$$\begin{aligned} throughput &= arrival\_rate \;-\; overflow\_rate \;-\; error\_rate \\ &= \lambda \;-\; \lambda P_{of} \;-\; \overline{P_p}(\lambda - \lambda P_{of}) \\ &= \lambda(1 - P_{of})(1 - \overline{P_p}). \end{aligned} \tag{3.5}$$

We consider the following optimization problem:

Throughput Maximization Problem*: At the beginning of each time slot* i*, given that the system state* $\boldsymbol{S}_i$ *is completely known by the transmitter and the receiver, select the transmission parameters* (U_i, P_i) *so that the system throughput is maximized, subject to an average transmission power constraint* $\overline{P}$*.*

3.3 Satisfying a BER Constraint

In this section, let us assume an extra constraint on our adaptive transmission scheme, that is the control action (U_i, P_i) must be selected so that a fixed BER constraint of $\overline{P_b}$ is satisfied. From a practical point of view, many existing communication protocols require that transmission must be carried out subject to a BER constraint. Furthermore, enforcing a BER constraint enables us to have a good comparison between our optimal adaptive transmission policies and those obtained in [GC97, CC99, BG02], where a BER constraint is also enforced.

Let $P(u, g, \overline{P_b})$ be the power needed to transmit u packets in a time slot of length T_s seconds when the channel state is g and the BER constraint is $\overline{P_b}$, and note that $P(u, g, \overline{P_b})$ depends on the specific coding, modulation, and detection schemes being used. For example, if an adaptive MQAM scheme as described in Section 3.2.3 is employed, and supposing that a transmission rate of u packets/slot is equivalent to mapping u bits to each modulated symbol, then from (3.4) we can approximate $P(u, g, \overline{P_b})$ by

$$P(u, g, \overline{P_b}) = \frac{WN_o}{\gamma_g}\left(\frac{-\log(5\overline{P_b})(2^u - 1)}{1.5}\right). \tag{3.6}$$

In this work, we assume that $P(u, g, \overline{P_b})$ has the general form of

$$P(u, g, \overline{P_b}) = \frac{WN_o}{\gamma_g} f(u, \overline{P_b}), \tag{3.7}$$

where $f(u, \overline{P_b})$ is increasing in u and decreasing in $\overline{P_b}$. $f(u, \overline{P_b})$ can be thought of as the received signal to noise ratio (SNR) needed to guarantee the BER of $\overline{P_b}$.

It should be noted that with the BER constraint, choosing a control action for time slot i, i.e., (U_i, P_i), is equivalent to choosing the transmission rate U_i, as after that, the transmit power P_i can be readily calculated.

As the BER performance is always kept at $\overline{P_b}$, from (3.2), the packet error probability $\overline{P_p}$ is always kept unchanged. When both λ and $\overline{P_p}$ are fixed, from (3.5), it is clear that maximizing the system throughput is equivalent to minimizing P_{of}. So from now on, we concentrate on minimizing the rate at which packets are dropped due to buffer overflow.

At the beginning of slot i, given that there are b packets in the buffer and we decide to transmit at rate u packets/slot, $u \in \{0, 1, \dots b\}$, the expected number of packets that are dropped at the end of slot i due to buffer overflow is

$$L_o(b, u) = \mathbb{E}\big\{ \max\{0, A + b - u - B\}\big\} \tag{3.8}$$

where the expectation is with respect to the distribution of A, i.e., the number of packets arriving in the slot. For example, if the arrival process is Poisson distributed with rate λ packets/second,

$$p_A(a) = \frac{\exp(-\lambda T_s)(\lambda T_s)^a}{a!} \tag{3.9}$$

and

$$L_o(b, u) = (\lambda T_s)\left(1 - \sum_{k=0}^{B-b+u-1} p_A(k)\right) - (B - b + u)\left(1 - \sum_{k=0}^{B-b+u} p_A(k)\right). \tag{3.10}$$

Our optimization problem can be written as:

$$\min_{U_0,\dots,U_{T-1}} \quad \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{i=0}^{T-1} L_o(B_i, U_i)\right\} \tag{3.11}$$

subject to:

$$U_i \in \{0, 1, \dots B_i\} \quad \forall i = 0, 1, \dots T - 1, \tag{3.12}$$

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{i=0}^{T-1} P(U_i, G_i, \overline{P_b})\right\} \leq \overline{P}. \tag{3.13}$$

Here, we assume that the set $\boldsymbol{\mathcal{P}}$ contains all power levels $P(u, g, \overline{P_b})$ for all $u \in \{0, \ldots B\}$ and all $g \in \{0, \ldots K-1\}$.

3.3.1 Optimal Policies (with a BER Constraint)

Instead of directly solving the above problem of minimizing the rate at which packets are dropped due to buffer overflow given an average power constraint, let us reformulate it as a problem of minimizing a weighted sum of the long term packet drop rate and average transmit power. In particular, we will find an adaptive transmission policy that minimizes

$$J_{avr} = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{i=0}^{T-1} C_I(B_i, G_i, U_i)\right\} \tag{3.14}$$

where $C_I(b, g, u)$ is the immediate cost incurred in state (b, g) when the transmission rate is set at u packets/slot, i.e.,

$$C_I(b, g, u) = P(u, g, \overline{P_b}) + \beta L_o(b, u). \tag{3.15}$$

In (3.15), β is a positive weighting factor that gives the priority to reducing packet loss over conserving power. In particular, by increasing β, we tend to transmit at a higher rate in order to lower the packet loss rate at the expense of more power being used. On the other hand, for smaller values of β, the average transmission power will be reduced at the cost of increasing packet loss rate. As it is pointed out in [BG02], if P^β and L^β are the average power and packet loss rate (due to buffer overflow) obtained when minimizing J_{avr} for a particular value of β, then L^β is also the minimum achievable loss rate subject to the average power constraint of P^β. In other words, for each value of β, minimizing J_{avr} gives us a Pareto optimal point (L^β, P^β) in the *Loss_Rate versus Power_Constraint* curve.

The problem of minimizing J_{avr} is an infinite horizon average cost Markov decision process (MDP) with system state $\boldsymbol{S}_i = (B_i, G_i)$, control action U_i, and immediate cost function $C_I(B_i, G_i, U_i)$. For an MDP to be completely defined, we also need to characterize the dynamics of the system when different control actions are selected. Supposing that the system state at time slot i is $\boldsymbol{S}_i = \boldsymbol{s} = (b, g)$ and a control action u is taken, the probability of the system being in state $\boldsymbol{s}' = (b', g')$ in the next time slot is

$$P_{\boldsymbol{S}}(\boldsymbol{s}, \boldsymbol{s}', u) = \Pr\{\boldsymbol{S}_{i+1} = \boldsymbol{s}' \mid \boldsymbol{S}_i = \boldsymbol{s}, U_i = u\} = P_G(g, g')P_B(b, b', u). \quad (3.16)$$

Here $P_G(g, g')$ is the probability of transitioning from channel state g into channel state g' and

$$P_B(b, b', u) = \Pr\{B_{i+1} = b' | B_i = b, U_i = u\}. \quad (3.17)$$

As all packets arriving to the buffer during slot i, i.e. A_i, are only added to the buffer at the end of this slot, we can write

$$B_{i+1} = q(B_i - U_i, A_i) \quad (3.18)$$

where

$$q(b, a) = \min\{b + a,\ B\}. \quad (3.19)$$

Based on (3.16), (3.17), (3.18), the system dynamics are well defined.

Let $\boldsymbol{\pi} = \{\mu_0, \mu_1, \mu_2, \ldots\}$ be a policy which maps system states into transmission rates for each slot i, i.e., $U_i = \mu_i(B_i, G_i)$. We would like to find a policy

$\boldsymbol{\pi}^*_{avr}$ that satisfies:

$$\begin{aligned}\boldsymbol{\pi}^*_{avr} &= \arg\min_{\boldsymbol{\pi}} J_{avr}(\boldsymbol{\pi}) \\ &= \arg\min_{\boldsymbol{\pi}} \left\{ \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left\{ \sum_{i=0}^{T-1} C_I(B_i, G_i, \mu_i(B_i, G_i)) \right\} \right\}.\end{aligned} \tag{3.20}$$

This problem can be solved efficiently using dynamic programming techniques [KV86]. Moreover, for our system in which all the states are connected, there exists a stationary policy $\boldsymbol{\pi}^*_{avr}$, i.e. $\mu_i \equiv \mu \ \forall \ i \in \mathbb{N}$, which is the solution to (3.20). Using a policy iteration algorithm, an optimal policy $\boldsymbol{\pi}^*_{avr}$ can be reached in a finite number of steps [KV86].

Finally, it is also useful to consider the discounted cost problem that finds:

$$\begin{aligned}\boldsymbol{\pi}^*_{\alpha} &= \arg\min_{\boldsymbol{\pi}} J_{\alpha}(b, g, \boldsymbol{\pi}) \\ &= \arg\min_{\boldsymbol{\pi}} \lim_{T\to\infty} \mathbb{E}\left\{ \sum_{i=0}^{T-1} \alpha^i C_I\left(B_i, G_i, \mu_i(B_i, G_i)\right) | B_0 = b, G_0 = g \right\},\end{aligned} \tag{3.21}$$

where $0 < \alpha < 1$ is the discounting factor and

$$J_{\alpha}(b, g, \boldsymbol{\pi}) = \lim_{T\to\infty} \mathbb{E}\left\{ \sum_{i=0}^{T-1} \alpha^i C_I\left(B_i, G_i, \mu_i(B_i, G_i)\right) | B_0 = b, G_0 = g \right\} \tag{3.22}$$

is the discounted cost given that the initial system state is (b, g) and policy $\boldsymbol{\pi} = \{\mu_0, \mu_1, \mu_2, \ldots\}$ is employed. As the immediate cost function C_I is bounded, the limit in (3.21) always exists. As shown in [KV86], when $\alpha \to 1$, the solution of the discounted cost problem, i.e., $\boldsymbol{\pi}^*_{\alpha}$, is stationary and converges to that of the average cost problem in (3.20). Moreover, let $J^*_{\alpha}(b, g)$ be the minimum discounted cost when $B_0 = b$ and $G_0 = g$, we have

$$\begin{aligned}J^*_{\alpha}(b, g) = \min_u \Big\{ & C_I\left(b, g, u\right) \\ & + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g, g') p_A(a) J^*_{\alpha}\left(q(b-u, a), g'\right) \Big\}.\end{aligned} \tag{3.23}$$

Equation (3.23) is particularly useful for analyzing the structure of the optimal policy.

3.3.2 Structure of Optimal Policies

In this section, we will point out an interesting structural characteristic of the optimal policy $\boldsymbol{\pi}^*_{avr}$ that satisfies (3.20). In particular, for certain FSMC models in which the fading process is correlated over time and when the transmission power constraint is relatively large, the optimal transmission power and rate are non-increasing in the channel gain. This is in contrast to the well known water-filling structure of the capacity-achieving link adaptation policy, which allocates more transmission power to good channel states and less power to bad channel states [GV97, GC97].

We will show the above effect for a simple FSMC model which has three possible states with channel gains: $0 = \gamma_0 < \gamma_1 < \gamma_2$. Moreover, we further restrict that channel state transitions after each time slot can only happen between adjacent states, i.e., $P_G(0,2) = P_G(2,0) = 0$ while $P_G(0,0), P_G(0,1), P_G(1,1), P_G(1,0)$, $P_G(1,2), P_G(2,2), P_G(2,1)$ are all positive.

We will study the structure of the policy $\boldsymbol{\pi}^*_\alpha$ that satisfies (3.21). As when $\alpha \to 1$, $\boldsymbol{\pi}^*_\alpha \to \boldsymbol{\pi}^*_{avr}$ [KV86], a structural property that is true for $\boldsymbol{\pi}^*_\alpha$, for all $\alpha \in (0,1)$, is also true for $\boldsymbol{\pi}^*_{avr}$.

Let us look at the insight behind the equation (3.23). When the system is in state (b,g), $b > 0, g > 0$, there are two effects of taking a control action u. By transmitting at rate u there is an immediate cost $C_I(b,g,u)$. However, the more we transmit in slot i, the fewer number of packets are left to the future stages. Therefore, the second effect of transmitting at rate u in state (b,g) is to reduce the future cost

$$C_F(b,g,u) = \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g,g') p_A(a) J^*_\alpha \left(q(b-u,a), g'\right). \tag{3.24}$$

For state (b, g) with $b > 0$, $g > 0$, let $0 \leq u_1 < u_2 \leq b$ be two possible transmission rates. We introduce the following notation:

$$\Delta_I(b, g, u_1, u_2) = C_I(b, g, u_2) - C_I(b, g, u_1) \tag{3.25}$$

and

$$\Delta_F(b, g, u_1, u_2) = C_F(b, g, u_1) - C_F(b, g, u_2). \tag{3.26}$$

Here, $\Delta_I(b, g, u_1, u_2)$ is the increase in immediate cost while $\Delta_F(b, g, u_1, u_2)$ is the reduction in future cost when the transmission rate is increased from u_1 to u_2. Clearly, action u_2 is more favorable than u_1 in state (b, g) if and only if $\Delta_I(b, g, u_1, u_2) < \Delta_F(b, g, u_1, u_2)$.

How the optimal transmission rate varies in g depends on how Δ_I, Δ_F vary in g. From (3.7), (3.15), and (3.25), we have

$$\begin{aligned} &\Delta_I(b, 1, u_1, u_2) - \Delta_I(b, 2, u_1, u_2) \\ &\qquad = W N_o \left(f(u_2, \overline{P_b}) - f(u_1, \overline{P_b}) \right) \left(\frac{1}{\gamma_1} - \frac{1}{\gamma_2} \right). \end{aligned} \tag{3.27}$$

We state the following lemma, of which a proof is given in Appendix A.

Lemma 3.3.1. *For each buffer state $b > 0$, there exists a constant β_o such that for every $\beta > \beta_o$ and $0 \leq u_1 < u_2 \leq b$, the following inequality holds:*

$$\Delta_I(b, 1, u_1, u_2) - \Delta_I(b, 2, u_1, u_2) < \Delta_F(b, 1, u_1, u_2) - \Delta_F(b, 2, u_1, u_2). \tag{3.28}$$

An intuitive explanation is as follows. As derived in (3.27), the left hand side of (3.28) does not depend on β. On the other hand, the right hand side of (3.28) depends on β. Suppose the power constraint is set to such a high value that the average cost is dominated by the buffer overflow cost at the outage state (state 0). Then, the effect of transmitting u packets at state (b, g) can be approximated by the reduction in the number of packets present in the buffer when the channel reaches outage some time later. As the fading process is correlated, when g increases, the average time to reach outage from state g also increases. Therefore, the future reward of transmitting u packets is decreasing in g. Moreover, as this effect depends on the buffer overflow cost, which is in turn scaled linearly by β, the rate at which Δ_F decreases in g can be made arbitrary large.

Given Lemma 3.3.1, we can prove the following structural characteristic of the optimal policy.

Theorem 3.3.2. *For each buffer state $b > 0$, let β_o be defined as in Lemma 3.3.1 and $\beta > \beta_o$, then the optimal transmission rate for each state (b, g), $g > 0$, is non-increasing in g.*

Proof. We present a proof by contradiction. Let u_1^* and u_2^* be the optimal transmission rate at states $(b, 1)$ and $(b, 2)$ respectively. Suppose $0 \leq u_1^* < u_2^* \leq b$. From (3.23) we have

$$C_I(b, 1, u_1^*) + C_F(b, 1, u_1^*) \;\leq\; C_I(b, 1, u_2^*) + C_F(b, 1, u_2^*). \tag{3.29}$$

Similarly

$$C_I(b, 2, u_2^*) + C_F(b, 2, u_2^*) \;\leq\; C_I(b, 2, u_1^*) + C_F(b, 2, u_1^*). \tag{3.30}$$

The inequality (3.29) implies:

$$\begin{aligned} \Delta_I(b, 1, u_1^*, u_2^*) &= C_I(b, 1, u_2^*) - C_I(b, 1, u_1^*) \\ &\geq\; C_F(b, 1, u_1^*) - C_F(b, 1, u_2^*) = \Delta_F(b, 1, u_1^*, u_2^*). \end{aligned} \tag{3.31}$$

Similarly, from (3.30):

$$\Delta_I(b, 2, u_1^*, u_2^*) \;\leq\; \Delta_F(b, 2, u_1^*, u_2^*). \tag{3.32}$$

From (3.31) and (3.32) we have:

$$\Delta_I(b, 1, u_1^*, u_2^*) - \Delta_I(b, 2, u_1^*, u_2^*) \;\geq\; \Delta_F(b, 1, u_1^*, u_2^*) - \Delta_F(b, 2, u_1^*, u_2^*), \tag{3.33}$$

which contradicts Lemma 3.3.1 and therefore, $u_1^* \geq u_2^*$. $\square$

Comment 1: Theorem 3.3.2 states that for a certain correlated fading channel model and average transmission power constraint, the optimal transmission rate is *non-increasing* in the channel gain. In fact, our numerical results show an even stronger effect, i.e., in some cases, the optimal transmission rate *decreases* when the channel gain increases. This is in sharp contrast to the water-filling structure of the conventional optimal power allocation policy over fading channels [GV97]. Please refer to Section 3.5 for more details.

Comment 2: In [BG02] a similar approach in characterizing the structure of the optimal transmission policy when the fading process is uncorrelated is given. Given that the fading is uncorrelated, inequality (3.28) is always false as its left hand side is always positive while the right hand side is always equal to zero. In that case, [BG02] shows that the optimal transmission rate is non-decreasing in channel gain.

3.4 Removing the BER Constraint

In the previous section, we assume that the transmit power and rate must be chosen so that a fixed BER constraint is satisfied. This assumption are also made in other related works such as [GC97, CC99, BG02]. However, there are two reasons for us to consider removing the BER constraint. First, by not imposing a BER constraint, it is possible to trade off between packet loss due to buffer overflow and packet loss due to transmission errors. For example, the transmitter can choose either to transmit at a high rate to reduce buffer overflow but suffering more transmission errors or to use a lower rate which leads to higher buffer overflow but at the same time reducing the transmission error probability. Second, it is not possible to meet a BER constraint when the transmitter only has an imperfect estimate of the channel gain. In that case, the transmitter can not calculate the transmit power and rate that guarantee any fixed BER constraint. This situation will be studied in Chapter 4.

3.4.1 Taking Transmission Errors into Account

For specific coding, modulation, and detection schemes being used, given the transmission rate u, power P, and channel state g, the bit error probability $P_b(g,u,P)$ can be estimated. Then (3.2) can be used to calculate the packet error probability $P_p(g,u,P)$. As a total of u packets are transmitted, the expected number of packets lost due to transmission error is

$$L_e(g,u,P) = uP_p(g,u,P). \tag{3.34}$$

When the packet arrival rate is fixed, maximizing the system throughput is equivalent to minimizing total packet loss rate due to both buffer overflow and transmission errors. So we have the following optimization problem:

$$\min_{U_0,\ldots,U_{T-1},P_0,\ldots,P_{T-1}} \quad \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{i=0}^{T-1}\left(L_o(B_i,U_i) + L_e(G_i,U_i,P_i)\right)\right\} \tag{3.35}$$

subject to:

$$U_i \in \{0, 1, \ldots B_i\} \quad \forall i = 0, 1, \ldots T-1, \tag{3.36}$$

$$P_i \in \boldsymbol{\mathcal{P}} \quad \forall i = 0, 1, \ldots T-1, \tag{3.37}$$

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{i=0}^{T-1} P_i\right\} \le \overline{P}. \tag{3.38}$$

3.4.2 Optimal Policies (without the BER Constraint)

Similar to the approach in Section 3.3.1, we can reformulate the constrained optimization problem in (3.35), (3.36), (3.37), (3.38) as a problem of minimizing a weighted sum of the total packet loss rate (due to buffer overflow and transmission error) and average transmission power. The only modification needed here is for the immediate cost function, which now becomes:

$$\widetilde{C}_I(b, g, u, P) = P + \beta\Big(L_o(b, u) + L_e(g, u, P)\Big). \tag{3.39}$$

At time i, let the system state be $\boldsymbol{S}_i = \boldsymbol{s} = (b, g)$ and a control action (u, P) is taken, the probability of the system being in state $\boldsymbol{s}' = (b', g')$ in the next time slot is still characterized by (3.16), (3.17), (3.18). An important point to note from (3.16), (3.17), (3.18) is that the chosen transmission power level P *does not have any effect* on the system dynamics. Therefore, given a choice of transmission rate u, the necessary and sufficient condition for a power level to be optimal is that it must satisfy

$$P = \arg\min_{P\in\boldsymbol{\mathcal{P}}} \widetilde{C}_I(b, g, u, P) = \arg\min_{P\in\boldsymbol{\mathcal{P}}} \{P + \beta L_e(g, u, P)\}. \tag{3.40}$$

In other words, in each system state, we only have to decide on which rate the transmitter should use. After that, the power level will follow directly by solving (3.40).

Let $\boldsymbol{\pi}$ be a stationary policy which maps system states into transmission rate for each slot i, i.e., $U_i = \boldsymbol{\pi}(B_i, G_i)$. Define

$$C_I^*(b, g, u) = \min_{P \in \boldsymbol{\mathcal{P}}} \widetilde{C}_I(b, g, u, P) \tag{3.41}$$

and

$$\widetilde{J}_{avr}(\boldsymbol{\pi}) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E} \left\{ \sum_{i=0}^{T-1} C_I^*(B_i, G_i, \boldsymbol{\pi}(B_i, G_i)) \right\}. \tag{3.42}$$

We have to solve for

$$\boldsymbol{\pi}^* = \arg\min_{\boldsymbol{\pi}} \widetilde{J}_{avr}(\boldsymbol{\pi}). \tag{3.43}$$

Again, this problem can be solved using dynamic programming techniques.

3.5 Numerical Results and Discussion

We have studied the problem of buffer and channel adaptive transmission under two scenarios, when transmission must be carried out so that a target BER is met and when transmission at flexible BER levels is allowed. In this section, the structure and performance of the adaptive policies obtained will be studied numerically. The performance criteria that we are interested in is the long term packet loss rate, due to either buffer overflow or transmission errors, per time slot. Also, we use the term *normalized packet loss rate* to refer to the packet loss rate that is normalized by the arrival rate λ.

3.5.1 System Parameters

The system for our numerical study is as follows. Packets arrive to the buffer according to a Poisson distribution with average rate $\lambda = 10^3$ and 3×10^3 packets/second. All packets have the same length of $L = 100$ bits. The buffer

Table 3.1: Channel states and transition probabilities (an 8-state FSMC obtained by quantizing a Rayleigh fading channel with average gain 0.8 and Doppler frequency 10 Hz).

State k	0	1	2	3	4	5	6	7
γ_k	0	0.1068	0.2301	0.3760	0.5545	0.7847	1.1090	1.6636
P_{kk}	0.9359	0.8552	0.8334	0.8306	0.8420	0.8665	0.9048	0.9639
$P_{k,k+1}$	0.0641	0.0807	0.0859	0.0835	0.0745	0.0590	0.0361	n.a.
$P_{k,k-1}$	n.a.	0.0641	0.0807	0.0859	0.0835	0.0745	0.0590	0.0361
p_k	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

length is $B = 15$ packets. The channel bandwidth is $W = 100$ kHz and the power density of AWGN noise is $N_o/2 = 10^{-5}$ Watt/Hz. We consider both cases of correlated and i.i.d. fading channels. For the correlated channel model, we use an 8-state FSMC as described in Table 3.1. This channel model is obtained by quantizing the fading range of a Rayleigh fading channel that has the average power gain $\overline{\gamma} = 0.8$ and Doppler frequency $f_D = 10$ Hz. Note that the value of Doppler frequency f_D in our study corresponds to users moving at a slow speed. For example, if the carrier frequency is around 1GHz, then $f_D = 10$ Hz corresponds to a movement at the speed of 3 meters/second. For the i.i.d. channel model, the values of the channel gains are the same as in Table 3.1, however, the channel evolves independently over time with all state transitions equiprobable.

Adaptive transmission is based on a variable-rate variable-power M-ary quadrature amplitude modulation (MQAM) scheme similar to that described in [GC97]. Let T_{sym} be the symbol period of the MQAM modulator and assume a Nyquist signaling pulse, $\text{sinc}(t/T_{sym})$, is used so that the value of T_{sym} is fixed at $1/W$ seconds. When the symbol period T_{sym} is kept unchanged, varying the signal constellation size of the modulator gives us different data transmission rates. As has been specified in Section 3.2, the power and rate adaptation are carried out in a slot-by-slot basis. Each time slot is F modulated symbol long and therefore, $T_s = FT_{sym}$. Here we set $F = L = 100$ so that when a signal constellation of size $M = 2^u$ is used, exactly u packets are transmitted from the buffer during each time slot.

As has been discussed in Sections 3.3 and 3.4, we consider two classes of buffer and channel adaptive transmission policies. In the first class of adaptive policies, transmit power and rate are selected subject to a BER constraint. We use (3.6) to approximate the power needed to transmit u bits per QAM symbol when the channel gain is γ_k and the BER constraint is $\overline{P_b}$. This class of policies is called *MDP_I*, i.e., MDP class I. The other class of adaptive policies is called *MDP_II.* In MDP_II policies, the BER constraint is removed and packet loss due to transmission errors is taken into account in the optimization, as described in Section 3.4. Also, for MDP_II policies, we assume that the set $\boldsymbol{\mathcal{P}}$ of possible power levels is finite. This makes it easier to solve for P from (3.40). Obviously, the more power levels we have, the better performance we would expect from the adaptive policy.

3.5.2 An Interesting Structural Property

First, let us look at the structure of MDP_I policies obtained by solving (3.20) for the correlated FSMC given in Table 3.1. These policies are found by the policy iteration algorithm given in [KV86]. In Fig. 3.2, we plot the optimal transmission rates of a MDP_I policy obtained when $\lambda = 10^3$ packets/sec, $B =$ 15 packets, $f_D = 10$ Hz, $\overline{P_b} = 10^{-3}$ and $\overline{P} = 16$dB. As can be seen, when the buffer occupancy is fixed, the optimal transmission rate increases when

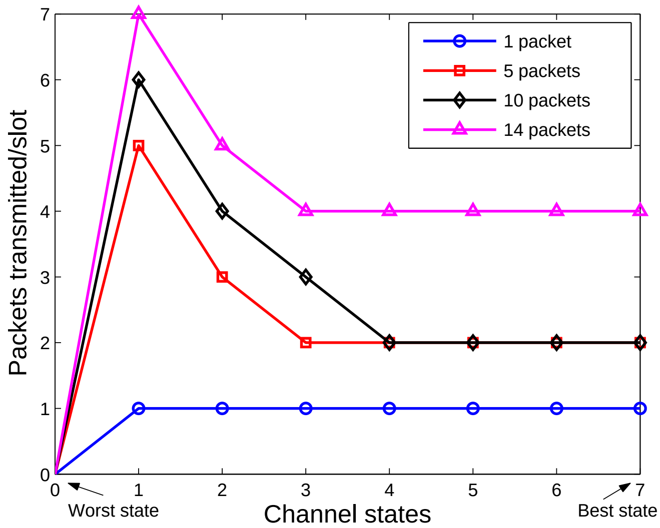

Figure 3.2: Structure of optimal policies, i.e., transmission rates (packets/slot) for different channel states when the buffer occupancy is fixed at 1, 5, 10 and 14 packets. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 1$ packet/slot, average power constraint $\overline{P} = 16$dB (the rest is given in Section 3.5.1). Channel model is correlated over time and given in Tab. 3.1. As can be seen, when the buffer occupancy is fixed, the transmission rate can increase when the channel gain decreases toward outage (state 0).

the channel gain decreases toward the outage point (state 0). This is consistent with our discussion in Section 3.3.2. For comparison, we also obtain the optimal policy for the i.i.d. channel model and plot its structure in Fig. 3.3. As can be seen, for each buffer occupancy, the optimal transmission rate increases when the channel gain increases. For numerous other sets of simulation parameters, similar effects have also been observed.

3.5.3 Packet Loss due to Buffer Overflow

Now we compare the performance of MDP_I policies with those of some other less adaptive schemes. All transmission is subject to a BER constraint of

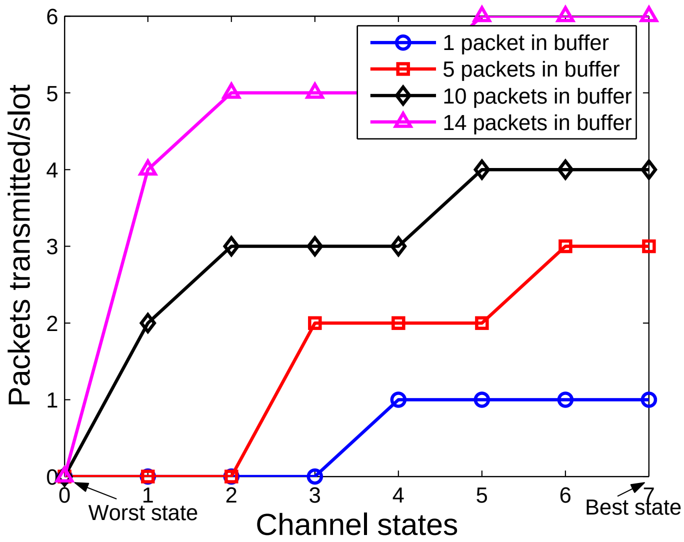

Figure 3.3: Structure of optimal policies, i.e., transmission rates (packets/slot) for different channel states when the buffer occupancy is fixed at 1, 5, 10 and 14 packets. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 1$ packet/slot, average power constraint $\overline{P} = 16$dB (the rest is given in Section 3.5.1). The fading process is i.i.d. over time. As can be seen, when the buffer occupancy is fixed, the transmission rate is non-increasing when the channel gain decreases toward outage (state 0).

$\overline{P_b} = 10^{-3}$ and we only care about packet loss due to buffer overflow. We consider two other classes of policies: channel inversion, i.e., *C_Inv*, and channel adaptive, i.e., *C_Adpt*. For each C_Inv policy, a fixed transmission rate is first selected. Based on this selected rate, the required SNR to meet the target BER is determined. Then, for each channel state with non-zeror gain, the transmit power is calculated based on inverting the channel gain to meet the required SNR. For channel state 0, i.e., when the channel gain is zero, the transmitter is turned off. In a C_Adpt policy, we use the optimal link-adaptive policy that maximizes the transmission rate for our channel model under some average power constraint and with the assumption that there are always packets to

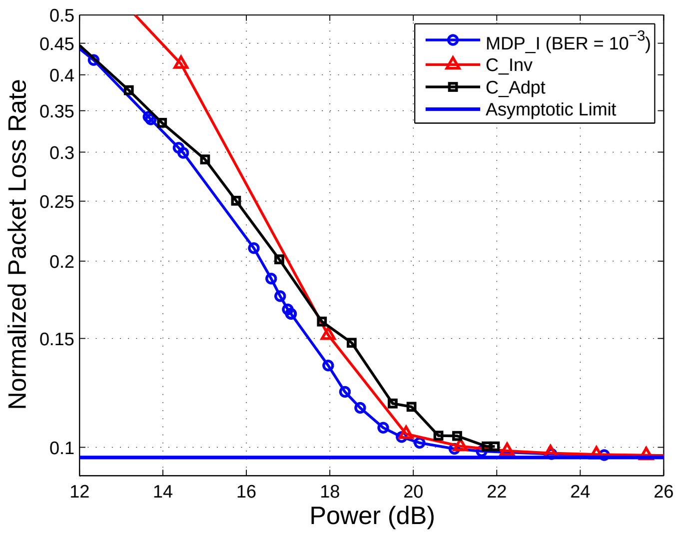

Figure 3.4: Performance, in terms of normalized packet loss rate (due to buffer overflow only) versus average transmission power, for MDP_I, C_Inv, and C_Adpt policies. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 3$ packets/slot, BER constraint $\overline{P_b} = 10^{-3}$ (the rest is in Section 3.5.1). Channel model is given by Table 3.1.

transmit. This scheme is equivalent to the variable-rate variable-power adaptive MQAM proposed in [GC97] (see Chapter 2, Section 2.1 for more detail). The performance of the three classes of policies, in terms of normalized packet loss rate (due to buffer overflow) versus the average power consumption are plotted in Fig. 3.4.

As it is expected, MDP_I outperforms the other two classes of adaptive policies. For low values of average transmit power, the performance of MDP_I policies and C_Adpt policies are very close while that of the C_Inv policies is much worse. This is expected, since at low power, the structure of an MDP_I policy is similar to that of the C_Adpt, and by focusing on conserving power, the system performance is improved. At high power, the performance of MDP_I and C_Inv policies are close and it is interesting to see that, for the same average transmit power, a C_Inv policy can result in less packet loss rate relative to a C_Adpt policy. This means that at this high range of average transmission power, if we only adapt to the channel, the performance can be worse than not doing any rate adaptation at all. We have looked at the performance of MDP_I, C_Adpt, and C_Inv policies for different values of Doppler frequency and observed that the performance of all schemes get worse when the Doppler frequency decreases. However, the relative difference between performance of different classes of adaptive policies does not seem to depend much on this parameter. We have also obtained results for longer buffer capacity and lower data arrival rate and observed that the performance trends of all schemes remain unchanged.

It can be noted in Fig. 3.4 that the packet loss rates of all policies, MDP_I, C_Adpt, C_Inv, reach a floor when the transmit power is high enough. This floor is represented by the asymptotic limit in Fig. 3.4. When the power constraint is high enough, the transmitter will always empty the buffer except in state 0. In that case, the floor for the normalized packet loss rate can be calculated exactly by:

$$L_{floor} = \frac{1}{\lambda}\left\{\frac{1}{K}\frac{\sum_{n=1}^{\infty} P_G(0,0)^{n-1}(1-P_G(0,0))L_a(n+1)}{\sum_{n=1}^{\infty}(n+1)P_G(0,0)^{n-1}(1-P_G(0,0))} + \frac{K-1}{K}L_a(1)\right\}, \tag{3.44}$$

where $L_a(n)$ is the expected number of packets lost due to buffer overflow during an interval of n time slosts, given that the buffer is empty at the beginning of the interval and no packet is further transmitted. Similar to (3.10), $L_a(n)$ can be calculated by:

$$L_a(n) = (\lambda n T_s)\left(1 - \sum_{k=0}^{B-b+u-1} \frac{\exp(-\lambda n T_s)(\lambda n T_s)^k}{k!}\right) - (B-b+u)\left(1 - \sum_{k=0}^{B-b+u} \frac{\exp(-\lambda n T_s)(\lambda n T_s)^k}{k!}\right). \tag{3.45}$$

Note that for the channel models in which there is no outage state, i.e., all channel states have positive gain, the packet loss floor will be $L_a(1)$. This discussion also holds for the problem in Chapter 4.

3.5.4 Packet Loss due to Buffer Overflow and Transmission Errors

Now we take packet transmission errors into account and compare the performance, in terms of total normalized packet loss rate versus average transmission power consumed, of two classes of buffer and channel adaptive transmission policies, namely MDP_I and MDP_II. To have a fair comparison between MDP_I and MDP_II policies, we also take into account packet transmission error for MDP_I policies. In particular, for a MDP_I policy with BER constraint set to $\overline{P_b}$, the packet error probability is

$$P_p = \sum_{j=l}^{L} \binom{L}{j} \overline{P_b}^j \left(1 - \overline{P_b}\right)^{(L-j)}. \tag{3.46}$$

The results, in terms of normalized packet loss rate (due to both buffer overflow and transmission error) versus power consumed, are plotted in Figs. 3.5, 3.6, 3.7 and 3.8.

Fig. 3.5 is for the case of correlated channel model. We plot the performance of the MDP_I policies corresponding to BER values of $10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$ and a MDP_II policies that have 20 different power levels, selected evenly from 4 to

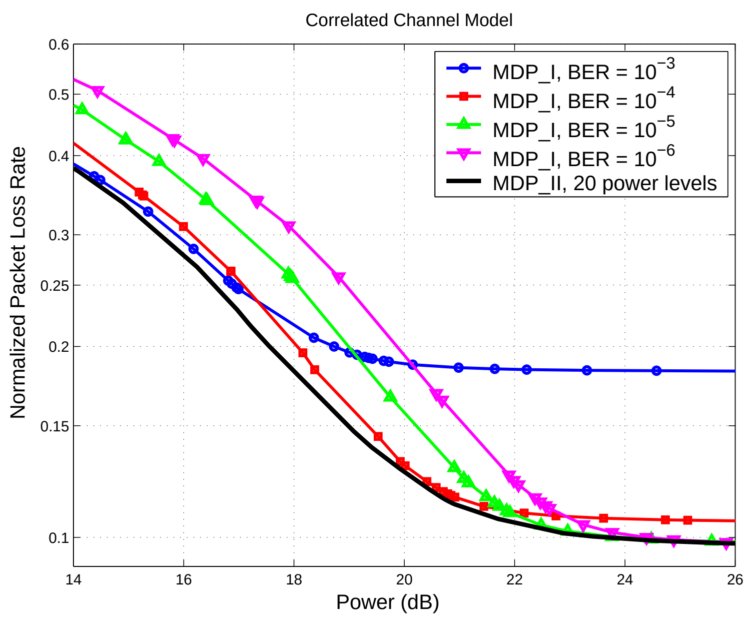

Figure 3.5: Performance, in terms of normalized packet loss rate (due to buffer overflow and transmission error) versus average transmission power, for MDP_I and MDP_II policies. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 3$ packets/slot, the rest is given in Section 3.5.1. Channel model is correlated over time and is given in Table 3.1.

40 dB. As can be seen, the MDP_II policies outperform MDP_I policies. MDP_I policies, corresponding to high values of BER, i.e. 10^{-3} and 10^{-4}, perform well in low ranges of transmission power while become much worse than the MDP_II policies in the high range of transmission power. On the other hand, for low value of BER, i.e. 10^{-6}, the performance of MDP_I policies is much worse than MDP_II policies in low power range. This can be explained by looking at the structure of the MDP_II policies. As an MDP_II policy can balance between packet loss due to buffer overflow and transmission errors, when the constrained power is low, it tends to transmit at relatively high BER values and when the constrained power is high, it transmits at low BER levels. In other words, at low power, the structure of a MDP_II policy is similar to those of MDP_I policies

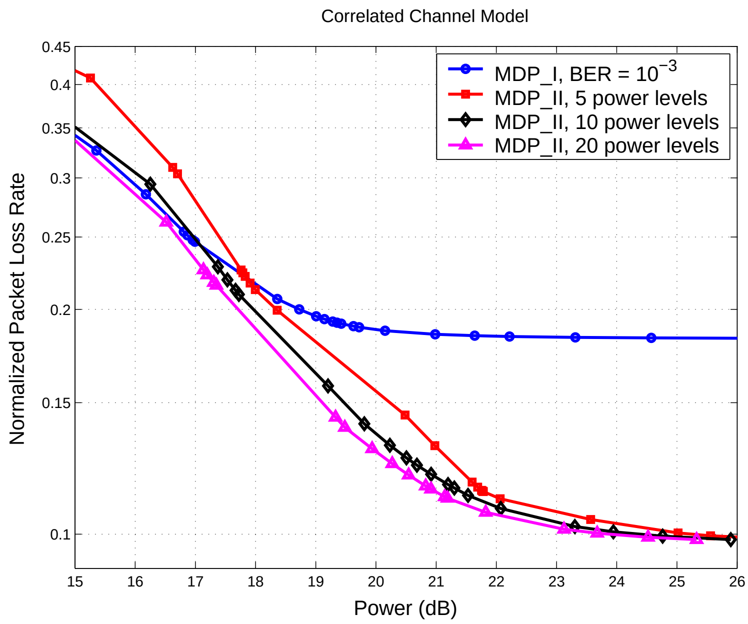

Figure 3.6: Performance, in terms of normalized packet loss rate (due to buffer overflow and transmission error) versus average transmission power, for MDP_I and MDP_II policies. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 3$ packets/slot, the rest is given in Section 3.5.1. Channel model is correlated over time and is given in Table 3.1.

corresponding to high BER constraints. On the other hand, when the power constraint is high, MDP_II policies are closer to MDP_I policies with low value of BER.

In Fig. 3.6, we plot the performance of different MDP_II policies that correspond to different numbers of possible power levels (from 4 to 40dB). As can be seen, even with only 5 different power levels, the MDP class II scheme can perform much better than MDP_I schemes.

Figs. 3.7 and 3.8 show result for i.i.d. channel models and similar effects can be observed.

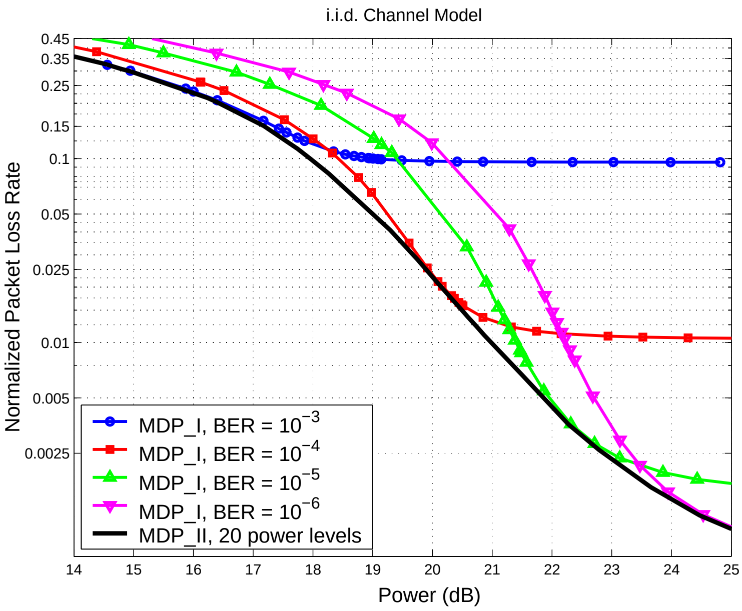

Figure 3.7: Performance, in terms of normalized packet loss rate (due to buffer overflow and transmission error) versus average transmission power, for MDP_I and MDP_II policies. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 3$ packets/slot, the rest is given in Section 3.5.1. Channel model is i.i.d. over time.

3.6 Conclusion

In this chapter, we considered the problem of buffer and channel adaptive transmission for maximizing the system throughput subject to an average transmit power constraint. Given that the instantaneous buffer and channel states are available for making control decisions, we reformulate the throughput maximization problem as a Markov decision process and solve for optimal transmission policies. Scenarios of incomplete system state information will be considered in Chapter 4.

This chapter highlights some important issues in wireless data communications. First, as nodes are only equipped with limited batteries and have to

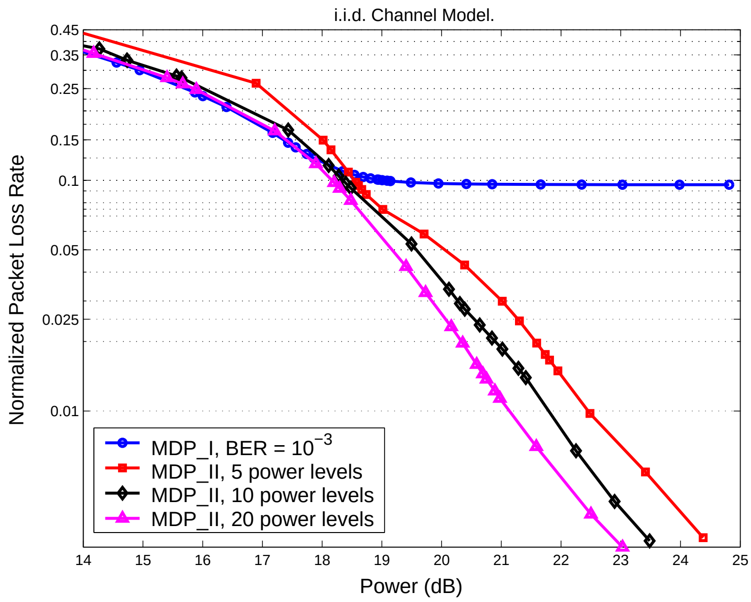

Figure 3.8: Performance, in terms of normalized packet loss rate (due to buffer overflow and transmission error) versus average transmission power, for MDP_I and MDP_II policies. System parameters are: buffer length $B = 15$ packets, arrival rate $\lambda = 3$ packets/slot, the rest is given in Section 3.5.1. Channel model is i.i.d. over time.

operate within a dynamic environment, i.e., with stochastic data arrivals and time-varying channels, cross-layer design approach is essential to achieve good performance. Second, when statistics at multiple layers are taken into account, popular intuitions associated with layered design can be no longer true. For example, our results show that the structure of the optimal buffer and channel adaptive transmission policies can be the reverse of water-filling.

CHAPTER 4

BUFFER AND CHANNEL ADAPTIVE TRANSMISSION: INCOMPLETE SYSTEM STATE INFORMATION

In this chapter, we continue studying the problem of buffer and channel adaptive transmission for maximizing the throughput of a single-user system. In Chapter 3, this problem has been studied under the assumption that the current system state, which consists of the instantaneous buffer occupancy and channel gain, is fully observable by both the transmitter and the receiver. Our focus in this chapter is for the cases when transmission decisions can only be based on some partial observation of the system state information (SSI).

In practice, for the SSI to be available at the transmitter and receiver, some processing and signaling are always required. For example, the transmitter can signal the receiver its buffer occupancy. At the same time, the channel state can be estimated by the receiver and then fed back to the transmitter. In Chapter 3, we assume that the process of estimating and signaling the channel and buffer states between the transmitter and the receiver is perfect so that control decisions can be made based on the exact knowledge of the current buffer occupancy and channel gain. In this chapter, we consider cases when the above processing and signaling are imperfect. In particular, the channel estimation/signaling process introduces delay and/or errors. At the same time, the buffer occupancy can be quantized to reduce the number of system states and lower the frequency of adapting the transmission mode.

Our focus is to study different buffer and channel adaptive transmission schemes that aim to maximize the system throughput given incomplete SSI. The main results of this chapter can be summarized as follows.

- We model different effects of incomplete SSI so that problem formulations fit into the framework of a partially observable Markov decision process (POMDP). Again, incomplete SSI includes quantized buffer occupancy and delayed and/or imperfectly estimated channel state.

- We discuss how buffer and channel adaptive transmission can be carried out given incomplete SSI. In particular, we show that optimal adaptive policies can be obtained for the cases when some delayed but error-free channel state information is available. When this is not possible, we propose various heuristics that achieve good performance.

- Finally, we present numerical results to support the theoretical development.

We note that some of the above results have been presented in [HM04a, HM05c].

As discussed in Chapter 3, Section 3.1, the problem we are considering is closely related to those studied in [GV97, GC97, CC99, SRB01, BG02, HGG02, GKS03, RSA04]. Our work shares the same objective with [GV97, GC97], i.e., to maximize the throughput of data transmission over a time-varying channel. In terms of cross-layer design, buffer and channel adaptive transmission schemes are also considered in [CC99, SRB01, BG02, HGG02, GKS03, RSA04].

This chapter focuses on cross-layer adaptive transmission under incomplete SSI. Works that study adaptive transmission under imperfect channel state information (CSI) include [GC97, Goe99, ZW02, OHH04, CG05]. Some of these works focus on characterizing the effects of channel estimation errors and delay on the spectral efficiency and BER performance of channel adaptive transmission schemes [GC97, OHH04]. Others explicitly incorporate imperfect channel estimates into the adaptation [Goe99, CG05, ZW02]. Especially, in [ZW02], Zhang and Wasserman consider channel adaptive transmission schemes given incomplete CSI and structure the problem as a POMDP. For more discussion on channel adaptive transmission under imperfect CSI, please refer to [CG05].

However, we note that there are few works considering cross-layer adaptive transmission under imperfect CSI. Two of the works following this line are [ZW02] (discussed above) and [HGG02]. In [HGG02], Holliday et al. consider the problem of buffer and channel adaptive transmission for minimizing the average transmit power subject to a constraint on the probability of packet loss due to deadline expiry. This work accounts for the effects of channel estimation errors and delay on the probability of packet retransmission.

The rest of this chapter is organized as follows. In Section 4.1, we discuss different situations in which the transmitter and receiver only have partial information about the current buffer and channel states. Section 4.2 is where we introduce two general approaches for buffer and channel adaptive transmission given incomplete SSI. One approach is MDP-based and the other is POMDP-based. In Section 4.3, we show that optimal control policies can be obtained when some delayed but error-free channel states are available for making decision. When this is not possible, we propose various heuristics to obtain policies with good performance in Section 4.4. Numerical results and discussion are given in Section 4.5. Finally, we conclude the chapter in Section 4.6.

4.1 Incomplete System State Information

In this section, let us discuss different scenarios in which only a partial observation of the current system state is available for making control decisions. In particular, the buffer occupancy can be quantized and the channel state can suffer from delay and/or errors.

4.1.1 Quantized Buffer State Information

Although the transmitter normally knows exactly what the current buffer occupancy is, we may not always want to adapt the transmission parameters to this exact value. The first reason is that the buffer occupancy can change frequently, therefore, adapting to its exact value may require a significant amount of signaling. Secondly, when the buffer length is long, the number of possible buffer states is large. This results in high complexity to find and implement the optimal buffer/channel adaptive policies.

We note that the need to quantize the buffer state will be even more important for multiple access scenario (as will be considered in Chapter 5). In that scenario, the base station may require the buffer information from all users in the system and by quantizing the buffer occupancies, the ammount of signalling and the size of the cntrol problem can be greatly reduced.

Due to the above reasons, we want to quantize the buffer occupancy using a small number of thresholds and only update the transmit power and rate when there is a threshold crossing. In this study, the buffer occupancy is quantized using $M+1$ thresholds, i.e., $0 = b_0 < b_1 < \ldots < b_M = L+1$. The buffer is said to be in state m, $m \in \{0, 1, \ldots M-1\}$, if the number of packets currently queueing satisfies $b_m \le b < b_{m+1}$. Denoting the quantized buffer occupancy at time i by B_i^{quant}, we have

$$B_i^{quant} = b_m, \text{ where } m \text{ satisfies } b_m \le B_i < b_{m+1} \ . \tag{4.1}$$

4.1.2 Delayed Error-free Channel State Information

We assume that the channel gain is first estimated at the receiver, then quantized into one of the possible values $\{\gamma_0, \gamma_1, \ldots \gamma_{K-1}\}$, and finally fed back to the transmitter. This process introduces both delay and errors in the transmitter knowledge's of the channel state. We discuss the delay factor first.

The delay in the channel state information available at the transmitter can be broken into estimation delay τ_e and feedback delay τ_f. In particular, τ_e is the processing time the receiver needs to obtain an estimate of the channel state while τ_f is the time to signal the estimate to the transmitter. Therefore, the estimated channel state is available at the transmitter after $\tau = \tau_e + \tau_f$ units of time.

In our model, as channel state transitions only happen at the beginning of each time slot, without loss of generality, we can assume that $\tau = mT_s$ where m is a non-negative integer. If we ignore the channel estimation errors for a moment and only concentrate on the effect of delay, then at the beginning of time slot i, the transmitter knows all channel states up to time slot $i-m$, i.e., $\{G_0, \ldots G_{i-m}\},\ i \geq m$.

4.1.3 Non-delayed Imperfect Channel Estimates

The channel state information available at the transmitter may suffer from estimation errors at the receiver and/or transmission errors on the feedback link. In this problem, we assume that a strong error correcting scheme is employed on the feedback link so that the feedback error is negligible. During time slot i, if we ignore the estimation/feedback delay, the sequence of imperfect channel estimates available at the transmitter can be denoted by $\{G_0^{est}, \ldots G_i^{est}\}$, where

G_i^{est} is an estimate of the channel state at time i. We account for the fact that G_i^{est} can be erroneous by the following function:

$$P_{ce}(g, \widehat{g}) = \Pr\{G_i^{est} = \widehat{g} \mid G_i = g\} \tag{4.2}$$

which gives the probability of wrongly estimating channel state g as channel state $\widehat{g}$. Note that $P_{ce}(g, \widehat{g})$ depends on the specific channel estimation technique employed at the receiver. In this study, we assume that the channel estimation error does not depend on the transmission parameters and is i.i.d. over time. We also assume that $P_{ce}(g, \widehat{g})$ is known at the transmitter for all pairs $(g,\ \widehat{g}),\ g,\ \widehat{g} \in \{0, \ldots K-1\}$.

As an example, let us assume that the estimation noise has a Gaussian distribution with zero mean and variance of σ^2, i.e., if the actual channel state is g, then the estimated channel gain prior to quantization is

$$\widehat{\gamma} = \gamma_g + n, \tag{4.3}$$

where n is a Gaussian random variable with zero mean and variance σ^2. The probability that $\widehat{\gamma}$ is closest to $\gamma_{\widehat{g}}$ can be written as

$$\begin{aligned} P_{ce}(g, \widehat{g}) = \frac{1}{2}\Big(\mathrm{erf}\Big(\frac{\gamma_{\widehat{g}} + \gamma_{\widehat{g}+1} - 2\gamma_g}{2\sqrt{2}\sigma}\Big) - \mathrm{erf}\Big(\frac{\gamma_{\widehat{g}} + \gamma_{\widehat{g}-1} - 2\gamma_g}{2\sqrt{2}\sigma}\Big)\Big), \\ 0 < \widehat{g} < K-1, \end{aligned} \tag{4.4}$$

and

$$P_{ce}(g, 0) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\gamma_0 + \gamma_1 - 2\gamma_g}{2\sqrt{2}\sigma}\right)\right), \tag{4.5}$$

$$P_{ce}(g, K-1) = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{\gamma_{K-2} + \gamma_{K-1} - 2\gamma_g}{2\sqrt{2}\sigma}\right)\right), \tag{4.6}$$

where erf(.) denotes the error function.

4.1.4 Delayed Imperfect Channel Estimates

If we take into account the effects of both delay and errors, then at time i, what available at the transmitter is a sequence of delayed imperfect estimates of the channel states up to time $i-m$, i.e., $\{G_0^{est}, \dots G_{i-m}^{est}\},\ i \geq m \geq 0$.

We also consider a special case in which the channel state information for choosing the transmit power and rate at time slot i is of the form $\{G_0, \dots G_{i-m-n}, G_{i-m-n+1}^{est}, \dots\ G_{i-m}^{est}\},\ i \geq m+n,\ m \geq 0, n \geq 0$. This means that, at time i, in addition to the imperfect channel estimates $\{G_{i-m-n+1}^{est}, \dots G_{i-m}^{est}\}$, the transmitter knows all the exact channel states up to time $i-m-n$. This assumption is justified by the fact that the accuracy of channel estimation can be improved if the receiver is given extra time and information to do processing [GC97]. For example, when a certain estimation delay is permitted, the receiver can interpolate between past and future estimates to obtain a more accurate prediction. Therefore, our assumption corresponds to the case when the delay $(m+n)T_s$ is long enough so that the receiver can obtain a near perfect channel estimate.

4.2 Adaptive Transmission under Incomplete SSI - General Approaches

In this section, we will discuss two main approaches to construct a buffer and channel adaptive transmission policy given incomplete SSI. One approach is based on the MDP solution to the problem when the system state is fully observable (see Chapter 3, Section 3.3 and 3.4). The other approach is based on formulating a partially observable MDP.

We note that, when the transmitter does not have the exact instantaneous channel state, it can not calculate the transmit power to guarantee a target BER at a given transmission rate. Therefore, in this chapter, we will only consider adaptive transmission policies that are not subject to a BER constraint (as in Chapter 3, Section 3.4).

4.2.1 Employing the MDP Policy $\boldsymbol{\pi}^*$

Perhaps the most straightforward approach is to employ the stationary buffer and channel adaptive policy $\boldsymbol{\pi}^*$ that minimizes a weighted sum of the long term packet loss rate and average transmit power when the current system state, i.e., buffer occupancy and channel gain, is completely known (see Chapter 3, Section 3.4). At time i, given a quantized buffer occupancy B_i^{quant}, and a channel estimate G_{i-m}^{est}, the chosen transmit power and rate are:

$$(U_i, P_i) = \boldsymbol{\pi}^*(B_i^{quant}, G_{i-m}^{est}). \tag{4.7}$$

We term this approach the *MDP approach.*

The MDP approach blindly assumes that the quantized buffer occupancy and/or estimated channel state are perfect. Later, we will introduce more complex approaches that account for the partial observability of the channel state. On the other hand, for quantized buffer occupancy, we will stick to this simple MDP approach. This is due to the following reasons. First, the probability distribution of the buffer occupancy depends on the adaptive transmission policy employed, therefore, it is highly complex to develop control schemes that account for the effects of buffer quantization. Second, unlike the case of incomplete channel state information, which originates from the limitations of the channel estimation and feedback process, quantization of the buffer occupancy is imposed to simplify the adaptive transmission policies. Therefore, we are not interested in developing complex algorithms that deal with the effect of buffer quantization.

For the rest of the chapter, when dealing with incomplete channel state information, we assume that the buffer state information is exact. When the buffer occupancy is indeed quantized, we will just use the quantized value directly in the place of the exact buffer occupancy.

4.2.2 Partially Observable MDPs

Instead of using the MDP policy $\boldsymbol{\pi}^*$, which blindly ignores the fact that the SSI available is incomplete, a more complex approach is to structure the problem as a partially observable Markov decision process (POMDP) and look for appropriate control policies. In addition to all components of an MDP, a POMDP model also specifies a stochastic observation process, i.e.,

$$P_O(x, o) = \Pr\{O_i = o \mid X_i = x\} \tag{4.8}$$

where X_i and O_i respectively denote the actual system state and its observation at time i. In our problem, the observations can be delayed and/or imperfectly estimated channel state.

In a POMDP, even though the underlying system is Markov, as the system state is only partially observed, the observation process may be non-Markovian. Therefore, the decision maker usually needs to keep track of some system memory or *internal system state* for choosing optimal control actions. Two popular choices for the internal system state are the *observation history* and the so called *belief state*. The observation history at time i is the sequence of all observations up to time i. Equivalently, a belief state which is a probability distribution over the set of all system states can be maintained during time slot i.

The main challenge in obtaining an optimal control policy for a POMDP is that the number of internal states is usually infinite. In that case, it is not possible to apply efficient dynamic programming algorithms as in a fully observable MDP. In our problem, when a delayed but error-free channel state can be obtained, the number of internal states is finite and an optimal control policy can be derived. For the cases when no error-free channel estimate can be obtained, the internal state space is indeed infinite and we can only approximate an optimal control policy. Details will be given in Sections 4.3 and 4.4.

4.3 Optimal Policies Given Delayed Error-free Channel States

We consider the special case, described in Section 4.1.4, when a combination of some delayed error-free channel states and less-delayed imperfect channel estimates is available for making a transmission decision in each time slot. In particular, the channel state information available at slot i is

$$\{G_0, \ldots G_{i-m-n},\ G^{est}_{i-m-n+1}, \ldots G^{est}_{i-m}\},\ i \geq m+n,\ m \geq 0, n \geq 0.$$

This means that the transmit power and rate for time slot i can be chosen based on all exact channel states up to time $i-m-n$ and a sequence of imperfect channel estimates $\{G^{est}_{i-m-n+1}, \ldots G^{est}_{i-m}\}$. The justification for this scenario is given in Section 4.1.4.

Due to the Markov property of the channel model, it is enough to only maintain a truncated sequence of the observation history, i.e., $\{G_{i-m-n}, G^{est}_{i-m-n+1} \ldots$

$G_{i-m}^{est}\}$. Now, the internal channel state at slot i can be defined as the vector

$$G_i^I = (G_{i-m-n}, G_{i-m-n+1}^{est} \ldots G_{i-m}^{est}). \tag{4.9}$$

As there are K possible channel states, the number of all possible internal channel states is K^{n+1}.

The important point to note is that even though the channel state information is incomplete, the number of internal states is still finite. This allows the problem of minimizing a weighted sum of the long term packet loss rate and average transmit power in Chapter 3, Section 3.4, to be formulated as a finite-state MDP, with the actual channel state G_i being replaced by the internal channel state G_i^I. In order to fully specify the MDP, we need to derive the dynamics of G_i^I, together with the cost functions associated with choosing transmission rate and power (u, P) in state (B_i, G_i^I). We will do this next.

4.3.1 Case When $m = 0,\ n = 1$

To simplify the derivations, we consider the case when $m = 0,\ n = 1$, i.e., at time i, the transmitter knows the exact previous channel state G_{i-1} and has an estimate of the current channel state G_i^{est}. The derivations for general values of m and n are similar.

When $m = 0,\ n = 1$, the internal channel state at time i is:

$$G_i^I = (G_{i-1},\ G_i^{est}). \tag{4.10}$$

The probability of transiting from state $G_i^I = (g_1^d, \widehat{g}_1)$ into state $G_{i+1}^I = (g_2^d, \widehat{g}_2)$

can be written as:

$$\begin{aligned}P_G^I((g_1^d,\widehat{g}_1),(g_2^d,\widehat{g}_2)) &= \Pr\{G_{i+1}^I=(g_2^d,\widehat{g}_2)\mid G_i^I=(g_1^d,\widehat{g}_1)\}\\ &= \Pr\{G_i=g_2^d, G_{i+1}^{est}=\widehat{g}_2)\mid G_{i-1}=g_1^d, G_i^{est}=\widehat{g}_1)\}\\ &= \boldsymbol{\mathcal{G}}(g_1^d,\widehat{g}_1,g_2^d)\sum_{g=0}^{K-1}P_{ce}(g,\widehat{g}_2)P_G(g_2^d,g).\end{aligned} \tag{4.11}$$

Given that $G_i^I=(g^d,\widehat{g})$, we can write down the probability distribution of the current channel state, i.e.,

$$\begin{aligned}\boldsymbol{\mathcal{G}}(g^d,\widehat{g},g) &= \Pr\{G_i=g\mid G_i^I=(g^d,\widehat{g})\}\\ &= \Pr\{G_i=g\mid G_{i-1}=g^d, G_i^{est}=\widehat{g}\}\\ &= \frac{P_{ce}(g,\widehat{g})P_G(g^d,g)}{\sum_{h=0}^{K-1}P_G(g^d,h)P_{ce}(h,\widehat{g})}.\end{aligned} \tag{4.12}$$

At time i, given a control action (u,P) when the buffer occupancy is $B_i=b$ and the internal channel state is $G_i^I=(g^d,\widehat{g})$, the average number of packets lost due to buffer overflow is still given by $L_o(b,u)$ while the expected number of packets lost due to transmission error is

$$L_e^I(G_i^I,u,P)=\sum_{g=0}^{K-1}\boldsymbol{\mathcal{G}}(g^d,\widehat{g},g)L_e(g,u,P). \tag{4.13}$$

Knowing the dynamics of G_i^I together with the cost of a transmission action in each state (B_i,G_i^I), an MDP can be readily formulated to minimize the weighted sum of the long term packet loss rate and average transmit power.

4.3.2 Case When $n=0$

Note that the number of all internal channel states is K^{n+1}. When $n=0$, i.e.,

$$G_i^I=G_{i-m}, \tag{4.14}$$

the number of internal channel states is K. As the number of possible internal channel states is the same as that of the actual channel states, the size of the newly formed MDP is the same as the size of the MDP for the case of complete channel state information (Chapter 3, Section 3.4).

4.4 Policies Given Imperfect Channel Estimates

In Section 4.3, we study cases when at each slot i, the transmitter knows exactly what the channel states up to time slot $i - m - n$ are. As discussed, when a delayed but error-free channel state is available at the transmitter, the number of internal states of the POMDP is finite and optimal control policy can be obtained. Now, we consider the general situation, described in Section 4.1.4, when a delayed error-free channel estimate is not available for choosing the transmit power and rate. In particular, at time i, we assume that the transmitter only knows a sequence of imperfect channel estimates $\{G_0^{est} \ldots G_{i-m}^{est}\}$.

4.4.1 Optimal Policies Given Delayed Imperfect Channel Estimates with i.i.d. Channel Model

In the special case when the channel state is independent, identically distributed over time, there is no extra information gained by keeping estimates of past channel states. We suppose that during time slot i, the transmitter knows the estimates of channel state G_i, i.e., G_i^{est}, then an internal channel state can be defined as

$$G_i^I = G_i^{est}. \tag{4.15}$$

The number of possible internal channel states is K and therefore, the problem of minimizing the weighted sum of packet loss rate and average transmit power can be formulated as a finite-state MDP. In particular, the dynamics of the internal channel state are easy to write down:

$$P_G^I(\widehat{g}_1, \widehat{g}_2) = \Pr\{G_{i+1}^{est} = \widehat{g}_2 \mid G_i^{est} = \widehat{g}_1\} = \sum_{g=0}^{K-1} P_{ce}(g, \widehat{g}_2) p_G(g). \tag{4.16}$$

Also, during time slot i, given that the channel estimate is $G_i^I = \widehat{g}$, we can derive the probability distribution of the current channel states as

$$\begin{aligned} \boldsymbol{\mathcal{G}}(\widehat{g}, g) &= \Pr\{G_i = g \mid G_i^I = \widehat{g}\} = \Pr\{G_i = g \mid G_i^{est} = \widehat{g}\} \\ &= \frac{P_{ce}(g, \widehat{g}) p_G(g)}{\sum_{j=0}^{K-1} P_{ce}(j, \widehat{g}) p_G(j)}. \end{aligned} \tag{4.17}$$

Given this distribution, all the cost functions can be derived.

4.4.2 Heuristic Policies Given Delayed Imperfect Channel Estimates

Now let us consider the case when the channel states are time-correlated. At time i, the transmitter only knows a sequence of delayed imperfect channel estimates $\{G_0^{est}, \ldots G_{i-m}^{est}\}$. To simplify the derivations, we further assume that $m = 0$. However, when $m > 0$ the analysis is similar.

When the channel is correlated over time, the decision maker needs to keep track of an entire channel estimation history, i.e., $\{G_0^{est}, \ldots G_i^{est}\}$, in order to select the optimal transmit power and rate. If we take the sequence $\{G_0^{est}, \ldots G_i^{est}\}$ as the internal channel state at time i, then the total number of internal channel states is infinite. Another option for the internal system state, which is more efficient to maintain, is the so called *belief channel state*. This is a K-element vector which specifies the probability distribution over K possible channel states at time i. In particular, let $\boldsymbol{\mathcal{G}}_i$ be the belief channel state at time i, then

$$\boldsymbol{\mathcal{G}}_i(g) = \Pr\{G_i = g \mid \boldsymbol{\mathcal{G}}_0,\ G_0^{est}, \ldots G_i^{est}\} \tag{4.18}$$

where the initial probability distribution $\boldsymbol{\mathcal{G}}_0$ is assumed known (in case $\boldsymbol{\mathcal{G}}_0$ is not given, it can be set to p_G, i.e., the stationary distribution of the channel states). The advantage of keeping a belief state for every time slot is that it contains all relevant information for making control actions. Furthermore, in the next time slot, given a new channel estimation $G_{i+1}^{est} = \widehat{g}$, the new belief state can be readily derived from

$$\boldsymbol{\mathcal{G}}_{i+1}(g) = \frac{P_{ce}(g, \widehat{g}) \sum_{g'=0}^{K-1} \boldsymbol{\mathcal{G}}_i(g') P_G(g', g)}{\sum_{g'=0}^{K-1} P_{ce}(g', \widehat{g}) \sum_{g''=0}^{K-1} \boldsymbol{\mathcal{G}}_i(g'') P_G(g'', g')}. \tag{4.19}$$

Unfortunately, maintaining a belief channel state for each time slot does not solve the problem of having infinite number of possible system states. When the number of system states is infinite, it is extremely hard to obtain an optimal adaptive policy. Doing so may require infinite time and memory. Therefore, instead of aiming for an optimal control policy, let us look at some approaches that can be used to approximate it. All of these approximations start with the assumption that we have already obtained the MDP policy $\boldsymbol{\pi}^*$ in (3.43), i.e., an optimal policy when the system state is fully observable.

Employing the MDP Policy $\boldsymbol{\pi}^*$

As discussed in Section 4.2.1, the most straightforward approach is to ignore the partial observability of the channel state and just employ policy $\boldsymbol{\pi}^*$, i.e., an optimal policy when the system state is fully observable. At time i, given the channel estimate G_i^{est} and buffer occupancy B_i, the transmission parameters are set as:

$$(U_i, P_i) = \boldsymbol{\pi}^*(B_i, G_i^{est}). \tag{4.20}$$

The Most Likely State Heuristic

In this approach, we first determine the state that the channel is most likely in, i.e.,

$$G_i^{MLS} = \underset{g \in \{0,\ldots K-1\}}{\arg\max} \; \{\boldsymbol{\mathcal{G}}_i(g)\} \tag{4.21}$$

Note that $\boldsymbol{\mathcal{G}}_i$ is the belief channel state at time i and is calculated using (4.19). Then the transmission parameters are set as:

$$(U_i, P_i) = \boldsymbol{\pi}^*(B_i, G_i^{MLS}). \tag{4.22}$$

This approach, which is usually termed the MLS approach, was proposed in [NPB95].

The QMDP Heuristic

This approach is related to the discounted cost problem defined in Chapter 3 (equation (3.21)). In particular, let the *Q function* be defined as:

$$Q(b, g, u, P) = C_I(b, g, u, P) + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g, g') p_A(a) J_\alpha^*(q(b-u, a), g'). \tag{4.23}$$

When the system state is fully observed, $Q(b, g, u, P)$ represents the cost of taking action (u, P) in state (b, g) and then acting optimally. The QMDP heuristic takes into account the belief state for one step and then assumes that the state is entirely known [LCK95]. In particular, the transmission rate and power for time i is chosen by:

$$(U_i, P_i) = \underset{u \in \{0,\ldots B_i\},\ P \in \boldsymbol{\mathcal{P}}}{\arg\min} \Big\{ \sum_{g=0}^{K-1} \boldsymbol{\mathcal{G}}_i(g) Q(B_i, g, u, P) \Big\}. \tag{4.24}$$

For more discussion on different approaches to approximate an optimal solution for a POMDP, please refer to [Lov91].

The Minimum Immediate Cost Heuristic

Finally, in order to assess the effectiveness of the MDP, MLS, and QMDP approaches, which are all MDP-based, we introduce a non-MDP heuristics which is called Minimum Immediate Cost (MIC). In MIC, at time slot i, given the belief state $\boldsymbol{\mathcal{G}}_i$, the transmission parameters are selected so that the expected immediate cost is minimized, i.e.,

$$(U_i, P_i) = \underset{u\in\{0,\ldots B_i\},\ p\in\boldsymbol{\mathcal{P}}}{\arg\min} \Big\{ \sum_{g=0}^{K-1} \boldsymbol{\mathcal{G}}_i(g) C_I(B_i, g, u, P) \Big\}. \tag{4.25}$$

4.5 Numerical Results and Discussion

4.5.1 System Parameters

The system for our numerical study is similar to that used in Chapter 3. Packets arrive to the buffer according to a Poisson distribution with average rate $\lambda = 3 \times 10^3$ packets/second. All packets have the same length of $L = 100$ bits. The buffer length is $B = 15$ packets. The channel bandwidth is $W = 100$ kHz and the power density of the AWGN noise is $N_o/2 = 10^{-5}$ Watt/Hz. We use two 8-state FSMCs as described in Tables 4.1 and 4.2. The channel model in Table 4.1 is obtained by quantizing the fading range of a Rayleigh fading channel that has average gain $\overline{\gamma} = 0.8$ and Doppler frequency $f_D = 10$ Hz while the one in Table 4.2 corresponds to $f_D = 20$ Hz.

Adaptive transmission is based on a variable-rate variable-power M-ary quadrature amplitude modulation (MQAM) scheme similar to that described in

Table 4.1: Channel states and transition probabilities (an 8-state FSMC obtained by quantizing a Rayleigh fading channel with average gain 0.8 and Doppler frequency 10 Hz).

State k	0	1	2	3	4	5	6	7
γ_k	0	0.1068	0.2301	0.3760	0.5545	0.7847	1.1090	1.6636
P_{kk}	0.9359	0.8552	0.8334	0.8306	0.8420	0.8665	0.9048	0.9639
$P_{k,k+1}$	0.0641	0.0807	0.0859	0.0835	0.0745	0.0590	0.0361	n.a.
$P_{k,k-1}$	n.a.	0.0641	0.0807	0.0859	0.0835	0.0745	0.0590	0.0361
p_k	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Table 4.2: Channel states and transition probabilities (an 8-state FSMC obtained by quantizing a Rayleigh fading channel with average gain 0.8 and Doppler frequency 20 Hz).

State k	0	1	2	3	4	5	6	7
γ_k	0	0.1068	0.2301	0.3760	0.5545	0.7847	1.1090	1.6636
P_{kk}	0.8718	0.7104	0.6668	0.6612	0.6841	0.7330	0.8097	0.9277
$P_{k,k+1}$	0.1282	0.1613	0.1718	0.1670	0.1489	0.1181	0.0723	n.a.
$P_{k,k-1}$	n.a.	0.1282	0.1613	0.1718	0.1670	0.1489	0.1181	0.0723
p_k	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

[GC97]. Let T_{sym} be the symbol period of the MQAM modulator and assume a Nyquist signaling pulse, $\text{sinc}(t/T_{sym})$, is used so that the value of T_{sym} is fixed at $1/W$ seconds. When the symbol period T_{sym} is kept unchanged, varying the signal constellation size of the modulator gives us different data transmission rates. The power and rate adaptation are carried out in a slot-by-slot basis. Each slot is F modulated symbol long and therefore, $T_s = FT_{sym}$. Here we set $F = L = 100$ so that when a signal constellation of size $M = 2^u$ is used, exactly u packets are transmitted from the buffer during each time slot.

Given a particular system state (b, g) and a control action (u, P), as derived in Chapter 3, the expected number of packets lost due to buffer overflow is

$$L_o(b,u) = (\lambda T_s)\left(1 - \sum_{k=0}^{B-b+u-1} p_A(k)\right) - (B-b+u)\left(1 - \sum_{k=0}^{B-b+u} p_A(k)\right) \quad (4.26)$$

where

$$p_A(a) = \frac{\exp(-\lambda T_s)(\lambda T_s)^a}{a!}. \quad (4.27)$$

We assume that a transmitted packet is in error if more than ten out of the 100 bits in the packet are in error. As in Chapter 3, the expected number of packets discarded due to transmission errors can be approximated by

$$L_e(g,u,P) = u\sum_{j=11}^{L}\left(\binom{L}{j}\left(P_b(g,u,P)\right)^j\left(1 - P_b(g,u,P)\right)^{(L-j)}\right) \quad (4.28)$$

where

$$P_b(g,u,P) = 0.2\exp\left(-1.5\frac{P\gamma_g}{WNo(2^u - 1)}\right). \quad (4.29)$$

We will look at the performance of different approaches discussed in Sections 4.2, 4.3, and 4.4 given incomplete SSI. When the packet arrival rate is fixed, maximizing the system throughput is equivalent to minimizing total packet loss due to buffer overflow and transmission error. Therefore, we will plot the long term packet loss rate versus average transmit power for each scheme.

4.5.2 Performance of MDP Policies Given Quantized Buffer Occupancy and Perfect Channel State

First, let us look at the performance of the MDP approach when the buffer occupancy is quantized. When the buffer occupancy is quantized, the performance of policy $\boldsymbol{\pi}^*$ depends on two factors, i.e., the number of quantized buffer states, and the selected quantization thresholds. Clearly, the more the number of quantized states, the closer the performance to the optimal. At the same

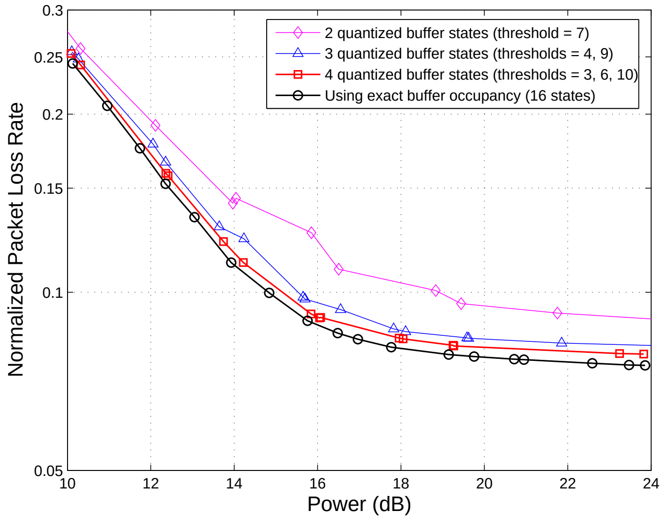

Figure 4.1: Performance of MDP $\boldsymbol{\pi}^*$ policy under quantized buffer state information. The performance is in terms of normalized packet loss rate versus average transmit power. System parameters are given in Section 4.5.1. Channel model is given in Table 4.2.

time, given a fixed number of quantized states, the performance depends on the set of selected thresholds. An intuitive way to select good quantization thresholds is to divide the range of buffer occupancy more finely at the range of high probability distribution. For example, if we know that most of the time, the buffer occupancy is low, then more thresholds should be set at low values.

In Fig. 4.1, we plot the performance of the MDP approach, in terms of total long term packet loss rate versus average transmit power, for different buffer quantization schemes. The number of quantized buffer states is increased from two to four. In particular, in the first quantization scheme, we set a single threshold at 7. When the buffer occupancy is less than 7, it is quantized to 0, otherwise, it is quantized to 7. Similarly, for the case of three quantized buffer states, we set the two thresholds at 4 and 9, and for the case of four quantized buffer states, we set the three thresholds at 3, 6, and 10. As can be seen, when only two quantized states are used, there is a significant loss compared to the case of adapting to the exact buffer occupancy. However, the packet loss rate is reduced significantly when the number of quantized buffer states is increased to three and four. When four quantized buffer states are used, the performance is quite near optimal. This suggests that we can quantize the buffer occupancy in order to reduce the complexity of the adaptive transmission policy without suffering significant performance degradation.

4.5.3 Performance of Different Approaches Given Delayed Error-free Channel State

Let us look at the performance of different buffer/channel adaptive transmission approaches when a delayed error-free channel state and an accurate buffer occupancy are available for making control decisions. We consider two scenarios. In the first scenario, at time slot i, the transmitter knows exactly what the channel state at time $i-1$, i.e., G_{i-1}, is. In the second scenario, in addition to knowing G_{i-1}, the transmitter also has an estimate of the channel state at time i, i.e., G_i^{est}. Both of these scenarios have been discussed in Section 4.3. In both cases, we have shown that optimal transmission policies, which maximize the system throughput given incomplete channel state information, can be obtained. To facilitate the discussion, we term the optimal adaptive policies under the first scenario the *POMDP_I* policies and the optimal adaptive policies under the second scenario the *POMDP_II* policies. Apart from these two classes of policies, we also look at the MDP approach which employed policy $\boldsymbol{\pi}^*$ (Section 4.2.1).

We plot the packet loss rate versus average transmit power for each approach.

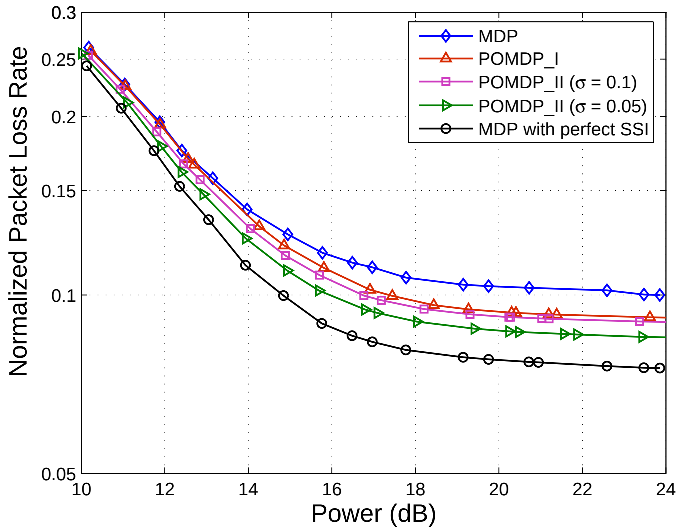

Figure 4.2: Performance, i.e., normalized packet loss rate versus average transmit power, for different adaptive transmission schemes given delayed error-free channel state information. Three schemes are considered, i.e., MDP (Section 4.2.1), POMDP_I, and POMDP_II (Section 4.3). System parameters are given in Section 4.5.1. Channel model is in Tab. 4.2.

Here, the packet loss rate is normalized by the average packet arrival rate. Clearly, the packet loss rates of all approaches are lower-bounded by the packet loss rate when optimal MDP policies are employed under perfect system state information. The performance of POMDP_I, POMDP_II, and MDP schemes are given in Figs. 4.2 and 4.3. Fig. 4.2 corresponds to channel model in Table 4.2 while Fig. 4.3 is for the channel model in Table 4.1.

In Figs. 4.2 and 4.3, we observe, as expected, that the performance of all schemes under delayed channel state information is lower-bounded by the performance of optimal transmission scheme with perfect channel knowledge. More importantly, the performance degradation increases when the channel changes faster (Fig. 4.2). This is expected because when the channel changes faster, the delayed channel state contains less information about the current

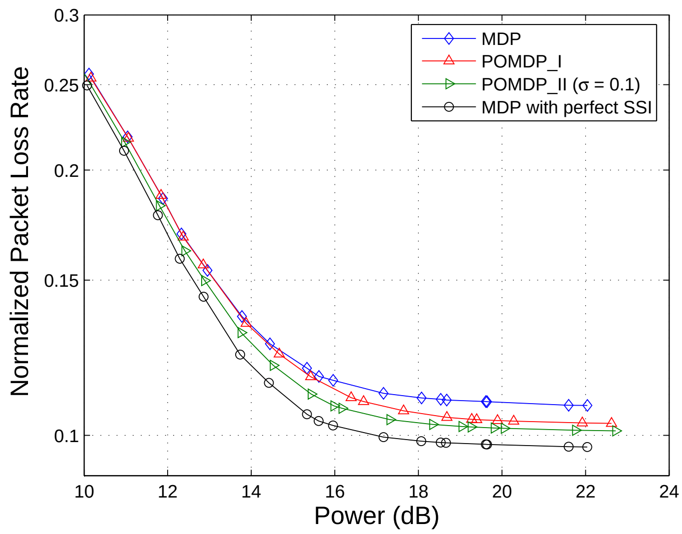

Figure 4.3: Performance, i.e., normalized packet loss rate versus average transmit power, for different adaptive transmission schemes given delayed channel state information. Three schemes are considered, i.e., MDP (Section 4.2.1), POMDP_I, and POMDP_II (Section 4.3). System parameters are given in Section 4.5.1. Channel model is in Tab. 4.1.

channel state.

The second observation that we can make from Figs. 4.2 and 4.3 is that the more information an adaptive scheme has, the better its performance is. In particular, POMDP_I policies perform better than MDP policies and POMDP_II policies perform better than POMDP_I. The performance of POMDP_II improves when the quality of the channel estimate G_I^{est} is improved. For example, when $\sigma = 0.05$, the performance of POMDP_II is quite close to that of the optimal scheme under perfect SSI. When the channel estimate G_i^{est} has high error probability ($\sigma = 0.1$), the performance of POMDP_II approaches that of POMDP_I. However, we note that the performance gain of POMDP_II comes at a cost of more complexity. In particular, the number of (internal) channel states for POMDP_II is K^2 while it is K for POMDP_I.

4.5.4 Performance of Different Approaches Given Imperfect Channel Estimates

Now let us look at the performance of different buffer and channel adaptive transmission approaches when no error-free channel state information is available at the transmitter. In particular, during time slot i, the transmitter only has a sequence of channel estimates $\{G_0^{est}, G_1^{est}, \ldots, G_i^{est}\}$. As has been discussed in Section 4.4.2, for the general case of correlated channel model, when no perfect channel estimate is available at the transmitter, it is not practical to look for optimal adaptive transmission policies. Instead, there are various heuristics that can approximate optimal control policies at lower complexity. These approaches are: MDP, MLS, QMDP and they have been discussed in Section 4.4.2. Again, we plot the performance of different adaptive approaches in terms of normalized packet loss rate versus average transmit power. The performance of all policies obtained are compared to the case when optimal MDP policies are employed under perfect SSI. The performance of different classes of adaptive policies is given in Figs. 4.4 and 4.5. Fig. 4.4 is obtained for the case when $\sigma = 0.05$ and Fig. 4.5 is for the case when $\sigma = 0.1$. In both Figs. 4.4 and 4.5, the channel model in Table 4.2 is used.

As can be seen, the MIC approach, which only tries to minimize the immediate cost during each time slot and does not take the dynamics of the system into account has the worst performance. Significant performance gain can be achieved by using MDP, MLS, and QMDP approaches. This shows the important of structuring the problem as a (partially observable) Markov decision process.

Among the three approaches MDP, MLS, and QMDP, it seems that QMDP

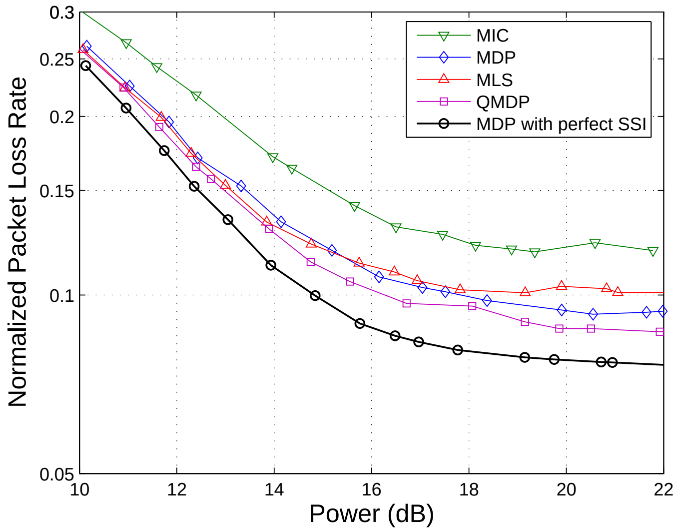

Figure 4.4: Performance, i.e., normalized packet loss rate versus average transmit power, for different adaptive transmission schemes given imperfect channel estimate. Three schemes are considered, i.e., MDP (Section 4.2.1), MLS (Section 4.4.2), QMDP (Section 4.4.2), and MIC (Section 4.4.2). System parameters are given in Section 4.5.1. Channel model is in Tab. 4.2. The standard deviation of channel estimating noise is $\sigma = 0.05$.

performs best. We note that there is no significant extra complexity when using QMDP instead of MDP or MLS, therefore, QMDP is a good choice to cope with imperfect estimated channel state information. Between MDP and MLS, MLS tends to perform better at low power range, while at higher power range, MDP achieves better results. However, we note that the difference in the performance of MDP and MLS is not significant, therefore, the simpler approach, i.e., MDP, is preferable.

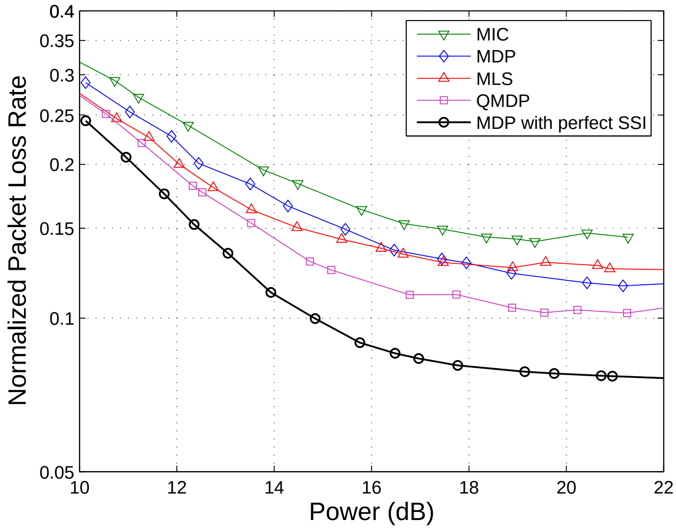

Figure 4.5: Performance, i.e., normalized packet loss rate versus average transmit power, for different adaptive transmission schemes given imperfect channel estimate. Three schemes are considered, i.e., MDP (Section 4.2.1), MLS (Section 4.4.2), QMDP (Section 4.4.2), and MIC (Section 4.4.2). System parameters are given in Section 4.5.1. Channel model is in Tab. 4.2. The standard deviation of channel estimating noise is $\sigma = 0.1$.

4.6 Conclusion

In this chapter, we considered the problem of buffer and channel adaptive transmission for maximizing the throughput of a transmission over a time-varying wireless channel, subject to an average transmit power constraint. We focused on scenarios in which the system state information for making control decisions is incomplete. This includes delayed and/or imperfectly estimated channel state and quantized buffer occupancy. We modeled the effects of partial observability so that they fit into the framework of a partially observable Markov decision process and showed how buffer and channel adaptive transmission can still be carried out.

Together with Chapter 3, this chapter shows the importance of cross-layer design in achieving good performance for wireless data communication system. The work presented in this chapter also demonstrates that, even when the system state is not fully observable, buffer and channel adaptive transmission can still be implemented in an effective manner. This means that our general approach of cross-layer adaptive transmission is robust with respect to uncertainty in knowledge of the system state.

CHAPTER 5

BUFFER AND CHANNEL ADAPTIVE SCHEDULING/TRANSMISSION FOR MULTIPLE-ACCESS WIRELESS CHANNELS

In Chapters 3 and 4, we have considered the problem of buffer and channel adaptive transmission for single-user systems. The results presented are also valid for multiple-access systems in which each user is assigned an orthogonal channel to transmit data. However, for those systems in which a channel is shared by multiple users, the results of Chapters 3 and 4 are not immediately applicable. In particular, before carrying out adaptive transmission, we need to consider how common channels are shared among users. This motivates us to study the problem of buffer and channel adaptive scheduling and transmission for multiple-access systems.

We consider a system in which multiple users transmit data packets to a base station over a time-varying wireless channel. Time is discretized into slots of equal length. During each time slot, data packets arrive to the buffers of transmitting nodes according to some stochastic distribution. All buffers are finite in length and packets arriving to a full buffer are lost. In each time slot, two controls decisions are made, i.e., a scheduling decision which assigns the common channel to one of the users and a transmission decision which sets the transmit power and rate for the scheduled user. All scheduling/transmission policies employed must satisfy an average transmit power constraint of each node.

Part of the objective of this chapter is to study optimal joint adaptive scheduling transmission policies that maximize the total system throughput.

Similar to the approach in Chapters 3 and 4, we obtain such an optimal policy by reformulating the throughput maximization problem as a Markov decision process (MDP). We note that when there are many users in the system, the complexity in obtaining and implementing the optimal joint scheduling/transmission policies can be very high.

A more important part of this chapter focuses on using the performance of optimal adaptive scheduling/transmission policies as a benchmark to assess other suboptimal adaptive scheduling/transmission policies that can be obtained and implemented at lower complexities. We start with two important classes of suboptimal policies: namely, *max-gain scheduling optimal transmission* and *round-robin scheduling optimal transmission*.

In max-gain scheduling optimal transmission policies, during each time slot, a user with the best channel condition is scheduled to transmit. Conditioned on this scheduling rule, the transmit power and rate of each user are selected based on the buffer and channel conditions so that the total system throughput is maximized. We show that the complexity of obtaining and implementing max-gain scheduling optimal transmission policies is significantly lower than that of optimal scheduling/transmission polices.

Similarly, in our round-robin scheduling optimal transmission policies, conditioned on a round-robin scheduling rule, the transmit power and rate of each user are selected so that the system throughput is maximized. We show that with the introduction of an *effective system state*, round-robin scheduling optimal transmission policies can be obtained at the same complexity as that of a single-user buffer/channel adaptive transmission problem.

Based on their performance, we identify the strength and weakness of each of the schemes: optimal, max-gain, and round-robin. We then propose hybrid schemes, which combine advantages of different schemes. Details of the hybrid schemes are given in Section 5.8.

In all adaptive scheduling/transmission schemes described above, it is assumed that the statistics of the data arrival processes and the time-varying channels are available for making control decisions. When this assumption is not satisfied, it is reasonable to consider adaptive schemes that are oblivious to these statistics. In [AKR$^+$01], Andrews et al. propose a scheme in which the user with the maximum product of the buffer occupancy and transmission rate is allowed to access the channel. We compare the performance of this scheme with the performance of those schemes described above.

The main results of this chapter can be summarized as follows.

- We formulate an optimization problem to find cross-layer adaptive scheduling/transmission policies that maximize the system throughput of a multiple-access system, subject to some average power constraints for all users.
- We show how MDPs can be formulated to obtain optimal as well as sub-optimal adaptive scheduling/transmission policies.
- By analyzing the performance and complexity of different class of adaptive scheduling/transmission policies, we come up with a design guideline, that can be used to determine the appropriate adaptive policy given a particular system setting.

We note that some results of this chapter has been presented in [HM04b].

The rest of this chapter is organized as follows. In Section 5.1, we discussed works that are related to the problem considered in this chapter. The problem of buffer and channel adaptive scheduling/transmission for maximizing the system throughput is described in Section 5.2. In Section 5.3, we show how optimal adaptive scheduling/transmission policies can be obtained. In Section 5.4, we briefly discuss a class of statistics-oblivious scheduling policies. This class of policies does not require the statistics of the data arrivals and channels. In Sections 5.5 and 5.6, max-gain scheduling optimal transmission policies and round-robin scheduling optimal transmission policies are respectively considered. The performance of these two classes of suboptimal policies are studied numerically in Section 5.7. Hybrid scheduling optimal transmission is discussed in Section 5.8. Finally, we conclude the chapter in Section 5.9.

5.1 Related Work

First, the problem considered in this chapter can be regarded as an extension to the works on cross-layer adaptive transmission over time-varying channels considered in Chapters 3, 4, and in [CC99, SRB01, BG02, HGG02, GKS03, RSA04] to a multiple-access scenario. In particular, these works consider the problem of adapting the transmit power and rate of a user to his buffer and channel conditions so that to minimize average queueing delay or maximize the system throughput. In our multiple-access setup, both the scheduling and transmission decisions are made based on the buffer and channel conditions of all users in the system.

The problem of maximizing the information theoretic capacity of a multiple-access system has been studied in a well-known work of Knopp and Humblet [KH95]. This work is discussed in details in Chapter 2, Section 2.2. The optimal scheduling/transmission scheme presented in [KH95] exhibits interesting properties: at each time instance, only a user who has the best channel condition is allowed to transmit and his power is allocated according to a water-filling strategy in time. However, always scheduling the user with the best channel condition can lead to unfairness among users. In [BW01, LCS01, LK03], channel adaptive scheduling policies that maximize the system throughput while satisfying different fairness constraints have been considered. The fairness constraints here can be in terms of the normalized throughput that each user has ([BW01, LK03]) or the fraction of resource each of them is assigned ([LCS01]). What makes our work different from [KH95] (and also from [BW01, LCS01, LK03]) is that we take into account the effects of a stochastic data arrival process and a limited buffer at each transmitter. In [KH95], it is implicitly assumed that an infinite amount of data is always available at each transmitter, therefore, the scheduled user can transmit at any rate that is determined by the water-filling power allocation algorithm.

Scheduling policies that take into account not only the channel conditions but also stochastic data arrivals and buffer occupancies have been considered in [TE93, AKR+01, SS02b, NMR03, LBH03, AKR+04]. The common objective of [TE93, AKR+01, SS02b, NMR03, AKR+04] is to find buffer and channel adaptive scheduling policies that guarantee the system stability, if there exists any scheduling policy that does so. In [TE93], for i.i.d. Bernoulli arrivals and i.i.d. on/off channel models, it is shown that the scheduling rule which serves the connected user with the longest queue makes the system stable. This result is generalized in [AKR+01, AKR+04], which state that two classes of policy called Modified Largest Weighted Delay First (M-LWDF) and Modified Largest Weighted Unfinished Work First (M-LWWF) achieve the system stability. In

[SS02b], another class of scheduling policies called the exponential rule is also shown to guarantee system stability. In [LBH03], the objective of optimizing the delay performance, which is captured by a utility function, is considered. Here, the utility function decreases in the delay experienced by packets in the buffers. The optimal policy that maximizes the system utility takes the packet delay and user transmission rate into account.

The works in [TE93, AKR+01, SS02b, NMR03, LBH03, AKR+04] focus mainly on the downlink scenario, i.e., from the base station to mobile terminals. On the other hand, our work is for the uplink scenario, i.e., from mobile terminals to the base station. It should be noted that the scheduling/transmission problems for uplink and downlink scenarios are different in many important aspects. Firstly, in the uplink scenario, the channel conditions of all users can be estimated by the base station. On the other hand the channel statistics must by estimated and fed back from individual users to the base station in the downlink scenario. Secondly, in the downlink scenario, all the buffers are at the base station, therefore, the buffer occupancies are readily available to the base station. On the other hand, these statistics must be sent by individual users to the base station in the uplink scenario. Finally, a fundamental difference is that in the downlink scenario, power conservation is not a pressing issue, as the base station is usually connected to a power supply. As a result, the works in [TE93, AKR+01, SS02b, NMR03, LBH03, AKR+04] focus only on the issue of scheduling multiple flows, and not on power and rate control.

Works that consider uplink scheduling or random access over wireless channels include [KQ03, VAT03, QB04]. In [KQ03], the problem of scheduling and power control for maximizing the weighted sum of users' achievable rates is con-

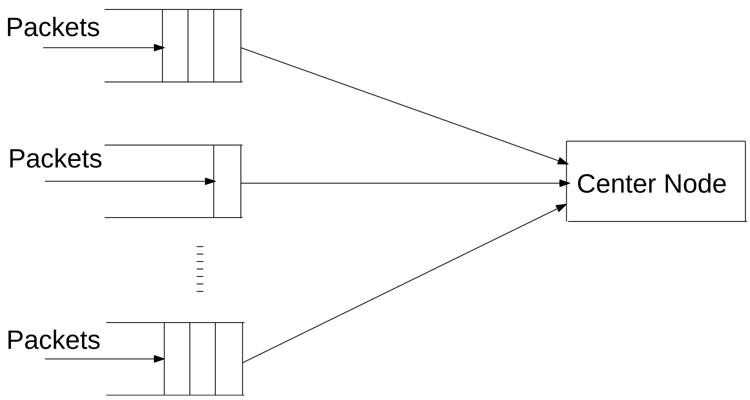

Figure 5.1: Model of a multiple-access data communication system.

sidered. However, the problem formulated in [KQ03] is independent for each time slot and does not take into account the fluctuations of wireless channels. The effect of packet loss due to finite-length buffers is also not of concerns in [KQ03]. The authors of [VAT03, QB04] propose different random access schemes that are based on slotted-ALOHA and aim to exploit the variations in channel conditions. The main contribution of the works in [VAT03, QB04] is to adapt the centralized capacity achieving scheme in [KH95] to distributed control scenarios.

5.2 Problem Description

5.2.1 System Model and General Notation

The system model considered in this chapter is depicted in Fig. 5.1. We have a discrete-time multiple-access system with stochastic data arrivals, finite-length buffers, and time-varying wireless channels. Important components of this system have been discussed in Chapter 2, Section 2.1. For the ease of following, we summarize the main assumptions and notations as follows.

- Time is discretized into slots of length T_s and time slot i, $i \in \mathbb{N}$, refers to the time period $[iT_s,\ (i+1)T_s)$.

- There are N users transmitting data packets to a base station via a time-varying wireless channel. N users are numbered from 1 to N and $\mathcal{N} = \{1, 2, \ldots N\}$ denotes the set of all users.

- During time slot i, there are A_i^n packets arriving to the buffer of user n, $n \in \mathcal{N}$. We assume that $\{A_i^n\}$ is stationary, ergodic, independent and identically distributed (i.i.d.) over time and across all users. Let λ and $p_A(a)$ denote the average and the stationary distribution of A_i^n respectively, λ and $p_A(a)$ are assumed known.

- Each user has a transmitter buffer of length B packets. B_i^n denotes the number of packets queuing in the buffer of user n at the beginning of time slot i. We assume that all packets arriving during time slot i are only added to the corresponding buffer at the end of the time slot. Packets that arrive when the corresponding buffer is full are dropped and considered lost.

- The time-varying channel conditions of N users are represented by N stationary, ergodic, finite state Markov channel (FSMC) models. We assume that the channel states stay constant during each time slot. Let G_i^n denote the channel state of user n during time slot i, we assume that $\{G_i^n\}$ is i.i.d. across all users and take one of the K possible states $\{0, 1 \ldots K-1\}$. Let g and g' be two possible channel states, i.e., $g,\ g' \in \{0, 1, \ldots K-1\}$, $p_G(g)$ denotes the steady-state probability of state g and $P_G(g, g')$ denotes the probability of transitioning from state g into state g' after each time slot.

- We define the system state in slot i as $\boldsymbol{S}_i = (B_i^1, B_i^2, \ldots B_i^N,\ G_i^1, G_i^2, \ldots G_i^N)$. Let $\mathcal{S}$ denote the set of all possible system states.

For more details on the data arrival processes, buffer dynamics, and finite-state Markov channels, please refer to Chapter 2.

Remark: It can be noted that we are considering a symmetric multiple-access system in which the parameters, i.e., data arrival statistics, buffer lengths, and channel statistics, are the same for all N users. This assumption is necessary for our proof of Theorem 5.3.1 in Section 5.3. Apart from that, the assumption of symmetric systems can be removed. The main reason for us to make it is for the sake of notational simplicity.

5.2.2 Cross-layer Adaptive Scheduling/Transmission Policies

We assume that the common channel is time-shared by N users, i.e., during each time slot, at most one user is allowed to transmit. Clearly, this assumption is valid for all systems that employ time division multiple access (TDMA). More generally, this assumption can be made in those systems in which the total bandwidth is first divided into a number of orthogonal channels and then each orthogonal channel is time-shared by a group of users. Note that the total bandwidth can be divided into independent channels using time, or frequency, or orthogonal code multiple access schemes. However, our time-shared assumption is not valid for code division multiple access (CDMA) systems that do not employ orthogonal codes.

When a user is assigned the channel, he can transmit data packets at different power levels and rates. We assume that all transmission must be subject to a reliability constraint which is transformed into the minimum power required to transmit at each particular rate. In particular, the power needed for each user to transmit reliably at rate u packets per time slot when his channel state is g, $g \in \{0, 1, \ldots K-1\}$, is denoted by $P(u, g)$. $P(u, g)$ is assumed to be non-negative and bounded for all u and g.

Let U_i^n denote the transmission rate of user n during time slot i. If user n is not scheduled during time slot i, then U_i^n is set to zero. Therefore, a scheduling/transmission decision in time slot i can be fully specified by the vector of transmission rate $\{U_i^1, U_i^2, \ldots U_i^N\}$. Based on this, we introduce the following definition of an adaptive scheduling/transmission policy.

Definition 5.2.1. *An adaptive scheduling/transmission policy is a sequence of functions* $\psi = \{\phi_0, \phi_1, \ldots\}$*, where for each time index* $i \in \mathbb{N}$*,* ϕ_i *is a map from* $\mathcal{S} \times \mathcal{N}$ *to the set of natural numbers* $\mathbb{N}$ *such that* $U_i^n = \phi_i(\boldsymbol{S}_i, n), \forall n \in \mathcal{N}$*. Furthermore,* ψ *is said to be* ***feasible*** *if and only if* $\forall\ i \in \mathbb{N}$*,*

$$\phi_i(\boldsymbol{S_i}, n)\phi_i(\boldsymbol{S_i}, m) = 0 \quad \forall m, n \in \mathcal{N},\ m \neq n, \tag{5.1}$$

and

$$\phi_i(\boldsymbol{S_i}, n) \ \leq \ B_i^n \quad \forall n \in \mathcal{N}. \tag{5.2}$$

In the above definition, (5.1) guarantees that at most one user can access the channel in each time slot while (5.2) means that the scheduled user cannot transmit more than what already available in his buffer. Note that in Definition 5.2.1, for different time slots, we allow different scheduling/transmission rules. We are also interested in stationary adaptive scheduling/transmission policies which are defined as

Definition 5.2.2. *An adaptive scheduling/transmission policy* $\psi = \{\phi_0, \phi_1, \ldots\}$ *is said to be **stationary** if and only if* $\phi_i = \phi_j, \ \forall i, j \ \in \ \mathbb{N}$.

Let $\boldsymbol{\Psi}$ be the set of all feasible adaptive scheduling/transmission policies. Also let $\boldsymbol{\Psi}^{st}, \ \boldsymbol{\Psi}^{st} \subset \boldsymbol{\Psi}$, be the set of all stationary feasible adaptive scheduling/transmission policies. Note that the cardinality of $\boldsymbol{\Psi}$ is infinite while that of $\boldsymbol{\Psi}^{st}$ is finite. In general, it is more convenient to deal with stationary policies. On the other hand, one advantage of non-stationary policies is that they can offer certain system fairness that can not be satisfied by any stationary scheduling policies. For example, when all users share the same buffer and channel conditions, a fair scheduling rule should give every user an equal chance to access the channel. This can be done in a non-stationary adaptive scheduling/transmission policy, but can not be satisfied by stationary scheduling policies.

5.2.3 Throughput Maximization Problem

We are interested in the following optimization problem.

Throughput Maximization Problem: *Find a feasible adaptive scheduling/transmission policy* $\psi \in \boldsymbol{\Psi}$ *that maximizes the system throughput, subject to the average power constraints of all* N *users. Here the system throughput is defined as the sum of the rates at which data packets are transmitted by all users in the system.*

Similar to our argument in Chapter 3 and 4, when the arrival rates are fixed for all users, maximizing the system throughput is equivalent to minimizing the total packet loss rate due to overflow at all the buffers. In time slot i, given the system state $\boldsymbol{S}_i = (B_i^1, B_i^2, \ldots B_i^N, \ G_i^1, G_i^2, \ \ldots G_i^N)$ and a scheduling/transmission decision $(U_i^1, U_i^2, \ldots U_i^N)$, the expected number of packets that are dropped due to buffer overflow for user $n \in \mathcal{N}$ is calculated by the function $L_o(B_i^n, U_i^n)$, where

$$L_o(B, u) = \mathbb{E}\big\{ \max\{0, A + b - u - B\}\big\} \tag{5.3}$$

The expectation in (5.3) is with respect to A - the number of packets arriving to the buffer of user n during the time slot of concern. Now, let $\overline{P}$ be the average transmit power constraint that each user must satisfies, our throughput maximization problem, or equivalently, the overflow minimization problem, can be stated as:

$$\min_{\psi \in \boldsymbol{\Psi}} \quad \limsup_{T \to \infty} \frac{1}{T} \mathbb{E} \left\{ \sum_{i=0}^{T-1} \sum_{n=1}^{N} L_o(B_i^n, \phi_i(\boldsymbol{S}_i, n)) \right\} \tag{5.4}$$

subject to:

$$\limsup_{T \to \infty} \frac{1}{T} \mathbb{E} \left\{ \sum_{i=0}^{T-1} P(\phi_i(\boldsymbol{S}_{\boldsymbol{i}}, n), G_i^n) \right\} \le \overline{P} \quad \forall n = 1, \ldots N. \tag{5.5}$$

Note again that in (5.4) and (5.5), $\psi = \{\phi_0, \phi_1, \ldots\}$.

5.3 Solving the Throughput Maximization Problem

5.3.1 Converting into a Non-constrained Optimization Problem

The problem of finding a throughput-maximizing adaptive scheduling/transmission policy in Section 5.2.3 consists of one objective function and N constraints. We study this problem using an approach similar to the one used in Chapters 3.

In particular, the constrained optimization problem is first reformulated into a non-constrained optimization problem of which the objective is to minimize the weighted sum of the total packet loss rate and total average transmit power, i.e.,

$$\min_{\psi \in \mathbf{\Psi}} \limsup_{T \to \infty} \frac{1}{T} \sum_{i=0}^{T} \sum_{n=1}^{N} \Big(\beta P(\phi_i(S_i, n), G_i^n) + L_o(B_i^n, \phi_i(S_i, n)) \Big), \tag{5.6}$$

where β is a weighting factor. Increasing β gives more priority to reducing transmit power, at the cost of more packet loss due to buffer overflow. On the other hand, reducing β puts more priority on reducing the packet loss rate, at the cost of more transmit power.

The problem in (5.6) can be regarded as an infinite horizon average cost Markov decision process (MDP). More importantly, as the weighted sum of total packet loss and total transmit power in each time slot is bounded and all the system states are connected, there exists a stationary policy $\psi^{\beta} \in \mathbf{\Psi}^{st}$ which is a solution to (5.6). Furthermore, the following theorem states the relationship between the performance of ψ^{β} and that of a policy solving the constrained throughput maximization problem in (5.4) and (5.5).

Theorem 5.3.1. *Let L_o^{β} and P^{β} be the average total packet loss rate and total transmit power corresponding to some stationary policy ψ^{β} that solves* (5.6)*, then L_o^{β} is also the minimum achievable total packet loss rate when each of the N users is subject to the average transmit power constraint of P^{β}/N.*

What stated by Theorem 5.3.1 is that, by solving (5.6) for a particular value of β, we obtain a pareto optimal point on the curve of *Packet_loss_rate versus*

Power_constraint. When β is varied, we can obtain different points on the optimal curve and that allows us to study the solution to the constrained throughput maximization problem in (5.4) and (5.5). Before proving Theorem 5.3.1, let us state the following lemma.

Lemma 5.3.2. *For any stationary feasible adaptive scheduling/transmission policy $\phi \in \mathbf{\Psi}^{st}$, let L_o^{ϕ} be the total packet loss rate of all users and P_n^{ϕ} be the average power consumed by user n when ϕ is employed, there exists a non-stationary policy $\psi \in \mathbf{\Psi}$ such that*

$$L_o^{\psi} \; = \; L_o^{\phi} \quad \text{while} \quad P_m^{\psi} \; = \; \frac{1}{N}\sum_{n=1}^{N} P_n^{\phi}, \quad \forall m \in \mathcal{N}, \tag{5.7}$$

where L_o^{ψ} is the total packet loss rate and P_m^{ψ} is the average power consumed by user m when policy ψ is employed.

A proof for Lemma 5.3.2 is given in Appendix B. Using this lemma, we present a proof for Theorem 5.3.1 as follows.

Proof. From Lemma 5.3.2, given stationary policy ϕ^{β} that solves (5.6) for a particular value of β, we can construct a non-stationary policy ψ^{β} that achieve the total packet loss rate of L_o^{β} while guaranteeing that the average power consumed by each user is P^{β}/N.

Now, suppose there exists another policy ψ, $\psi \in \mathbf{\Psi}$, that results in

$$L_o^{\psi} \; < \; L_o^{\beta}, \quad \text{while} \quad P_n^{\psi} \; \leq \; P^{\beta}/N. \tag{5.8}$$

This contradicts the assumption that ϕ^{β} is the solution of (5.6). Therefore, L_o^{β} is the minimum achievable packet loss rate when all users are subject to the average power constraint of P^{β}/N. $\square$

5.3.2 Markov Decision Process

The minimization problem in (5.6) can be regarded as an infinite horizon average cost Markov decision process (MDP). This MDP is specified by the following components.

System states: The system state at time i is $\boldsymbol{S}_i = (B_i^1, B_i^2, \ldots B_i^N,\ G_i^1, G_i^2, \ldots G_i^N)$ with $B_i^n \in \{1, 2, \ldots B\}$ and $G_i^n \in \{0, 1, \ldots K-1\}$. Note that the number of possible system states is finite.

Control actions: Given state $\boldsymbol{S}_i$ the control actions is specified by an N-element vector of integers $(U_i^1, U_i^2, \ldots U_i^N)$, where U_i^n is the transmission rate of user n in time slot i. Note that we must have:

$$U_i^n \in \{0, 1, \ldots B_i^n\} \quad \text{and} \quad U_i^n U_i^m = 0,\ \forall n, m \in \mathcal{N}, m \neq n. \tag{5.9}$$

Immediate cost function: The immediate cost of choosing action $(U_i^1, U_i^2, \ldots U_i^N)$ in state $\boldsymbol{S}_i$ is the sum of weighted total power consumed and total packet loss due to buffer overflow, i.e.,

$$C(\boldsymbol{S}_i, U_i^1, U_i^2, \ldots U_i^N) = \sum_{n=1}^{N} \Big(\beta^n P(U_i^n, G_i^n) + L_o(B_i^n, U_i^n)\Big). \tag{5.10}$$

System dynamics: To fully specify the MDP, we also need to characterize the system dynamics, i.e., the transitioning probability of the system states, given control actions. The system states consist of the buffer occupancies and channel states of all N users. We note that the channel transition probabilities are independent of the control actions. For user n, $n \in \mathcal{N}$, as all packets arriving to the buffer during frame i, i.e. A_i^n, are only added to the buffer at the end of this frame, we can write

$$B_{i+1}^n = \min\{B_i^n - U_i^n + A_i^n,\ B\} \tag{5.11}$$

Knowing the channel state transition probabilities, as well as the distribution of packet arrival processes, the dynamics of the system state can be readily characterized.

The infinite horizon, average cost MDP in (5.6) can be solved using dynamic programming techniques such as value iteration or policy iteration [KV86].

5.3.3 Complexity of Obtaining and Implementing Throughput Maximizing Policies

The difficulty in obtaining a solution to the optimization problem in (5.6) using dynamic programming techniques lies in the fact that the size of the system state space of the corresponding MDP can be very large. In particular, the total number of possible system states is $(B+1)^N K^N$. This results in a high complexity in solving the corresponding MDP when the number of users in the system and their buffer length and/or number of channel states increase.

As the operation of optimal adaptive scheduling/transmission policies requires knowledge of the buffer occupancies and channel states of all N users, it is more reasonable to implement these adaptive policies at the base station, rather than doing so at each individual node. The optimal policies can be stored at the base station using a look-up table. At the beginning of each time slot, all users signal their buffer occupancies to the base station. As for the channel states, they can be estimated by the base station. The base station then outputs a scheduling/transmission decision based on the current system state. We note that when there are many users in the system and their buffer lengths are long, the amount of signaling required to transmit the buffer occupancies to the base station can be significant.

5.4 Statistics-oblivious Adaptive Scheduling Policies

As has been discussed in Section 5.1, our work is related to the problem of scheduling parallel queues over downlink time-varying wireless channels [TE93, AKR+01, SS02b, NMR03, LBH03, AKR+04]. However, as power conservation is not an objective for downlink scheduling, these works do not consider adapting the transmit power and rates. Instead, it is assumed that there are some underlying mechanism at the physical layer to associate each channel state with a transmission rate and the focus is on designing adaptive scheduling policies that take buffer lengths and transmission rates into account.

An important result in the downlink scheduling problem is that there are some classes of adaptive scheduling policies, which are oblivious to the data arrival and channel statistics while still able to maintain the stability of the system (if there is any other scheduling policy that does so). An example is the policy which always assigns the channel to a user who has the maximum product of instantaneous buffer occupancy and transmission rate [AKR+01, AKR+04].

Inspired by the above result, we consider the following class of adaptive scheduling/transmission policies. For each channel state, associate a maximum rate at which data can be transmitted. This maximum transmission rate can be set by assuming that each user transmits at some fixed transmit power so that transmission rate can be readily calculated for each channel state. For example, let P^c be some chosen transmit power for all users, then, if user n is scheduled in time slot i, his maximum transmission rate is

$$R_i^n = \max\{r \in \mathbb{N} \mid P(r, G_i^n) \ \leq \ P^c\}. \tag{5.12}$$

Now, for time slot i, allow the user n who has the maximum $B_i^n R_i^n$ to transmit.

For the selected user, if there are enough data to transmit at the maximum rate, then the maximum rate is used. Otherwise, transmit all the data in the buffer. In a more concrete form, we define a max-product scheduling/transmission policy as follow.

Definition 5.4.1. *A* ***max-product scheduling adaptive transmission policy*** *is a feasible adaptive scheduling/transmission policy* $\psi = \{\phi_0, \phi_1, \ldots\}$, $\psi \in \boldsymbol{\Psi}$*, such that* $\forall i \in \mathbb{N}$, $\phi_i(\boldsymbol{S}_i, n) = 0$ *if:*

$$B_i^n R_i^n \; < \; \max_{m \in \mathcal{N}} \{B_i^m R_i^m\}. \tag{5.13}$$

Note that max-product policies are oblivious to the statistics of the data arrivals and channel fluctuations. All that are required are the instantaneous buffer occupancies and transmission rates of all users. We will study the performance of this class of policies in Section 5.7.

5.5 Max-gain Scheduling Optimal Transmission

We note that both optimal scheduling/transmission policies and max-product scheduling policies require the buffer and channel states of all N users. This makes these policies not suitable for implementing at each individual node. For the implementation at the base station, significant amount of signaling may be required to transmit the buffer occupancies to the base station. In this section and Section 5.6, we will look at some scenarios in which the scheduling rule is independently to the users' buffer conditions. This reduces the amount of signaling required when the adaptive policies are implemented at the base station and even makes it possible to implement them at each individual nodes.

5.5.1 Max-gain Scheduling Adaptive Transmission Policies

Let us consider an adaptive scheduling rule which, during each time slot, allows a user with the best channel condition to transmit. We term this *max-gain scheduling*. For the max-gain scheduling rule to be well-defined, we still need to specify how the channel is assigned when there are more than one user with the best channel condition. One way to do this is by assigning N distinct priority levels to N users and when there are more than one user with the best channel condition, the one with the highest priority level is scheduled. Another way (which is used here) is to select the users with equally best channel condition with equal probabilities. Formally, we define a max-gain scheduling adaptive transmission policy as follows.

Definition 5.5.1. *A **max-gain scheduling adaptive transmission policy** is a feasible adaptive scheduling/transmission policy* $\psi = \{\phi_0, \phi_1, \ldots\},\ \psi \in \mathbf{\Psi}$*, such that* $\forall i \in \mathbb{N},\ \phi_i(\boldsymbol{S}_i, n) = 0$ *if:*

$$G_i^n \ < \ \max_{m \in \mathcal{N}} \{G_i^m\}. \tag{5.14}$$

In the above definition, $\mathbf{\Psi}$ is the set of all feasible adaptive scheduling/transmission policies as defined in Definition 5.2.1.

Before moving on, we note that max-gain scheduling is inspired by the work in [KH95], which shows that the variation in the channel conditions across users introduces a form of multiuser diversity which can be optimally exploited by always allocating all available bandwidth to the user with the best channel condition. Now let $\mathbf{\Psi}^{mg}$ denote the set of all max-gain scheduling adaptive transmission policies, we would like to solve the following problem:

$$\min_{\psi\in\mathbf{\Psi}^{mg}} \limsup_{T\to\infty} \frac{1}{T}\sum_{i=0}^{T}\sum_{n=1}^{N}\Big(\beta P(\phi_i(S_i,n),G_i^n)+L_o(B_i^n,\phi_i(S_i,n))\Big). \qquad (5.15)$$

5.5.2 Obtaining Max-gain Scheduling Optimal Transmission Policies

We are going to show that (5.15) can be decomposed into N simpler optimization problems.

For user n, $n\in\mathcal{N}$, let us consider the following problem

$$\min_{\psi\in\mathbf{\Psi}^{mg}} \limsup_{T\to\infty} \frac{1}{T}\sum_{i=0}^{T}\Big(\beta P(\phi_i(S_i,n),G_i^n)+L_o(B_i^n,\phi_i(S_i,n))\Big). \qquad (5.16)$$

In (5.16), we would like to find a max-gain scheduling adaptive transmission policy that minimizes the weighted sum of the packet loss rate and average transmit power *for user n*. Due to its special structure, the problem in (5.16) can be reduced in size. In particular, let us define the reduced system state of user n in time slot i as

$$\boldsymbol{S}_i^{mg,n}=(B_i^n,G_i^1,G_i^2,\dots G_i^N). \qquad (5.17)$$

This reduced system state consists of the current buffer occupancy of user n, together with the current channel conditions of all N users.

Conditioned on the max-gain scheduling rule, let $\mathbf{\Pi}_n^{mg}$ be the set of all transmission policies μ which set the transmission rate U_i^n for user n based on $S_i^{mg,n}$, i.e., $U_i^n=\mu(S_i^{mg,n})$. Clearly, all policies $\mu\in\mathbf{\Pi}_n^{mg}$ must satisfy the following conditions in (5.18) and (5.19)

$$\mu(S_i^{mg,n})\in\{0,1,\dots B_i^n\}, \qquad (5.18)$$

$$\mu(S_i^{mg,n}) = 0 \text{ if } G_i^n < \max_{m \in \mathcal{N}} G_i^m. \tag{5.19}$$

Now, let $\mu^{n,*}$ be a (stationary) policy in $\mathbf{\Pi}_n^{mg}$ such that

$$\mu^{n,*} = \arg\min_{\mu \in \mathbf{\Pi}_n^{mg}} \limsup_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} \Big(\beta P(\mu(\boldsymbol{S}_i^{mg,n}), G_i^n) + L_o(B_i^n, \mu(\boldsymbol{S}_i^{mg,n})) \Big). \tag{5.20}$$

Using the technique of homogeneous immediate reward partition (introduced in [DG97]), it can be shown that any policy $\psi \in \mathbf{\Psi}$ that satisfies

$$\psi(\boldsymbol{S}_i, n) = \mu^{n,*}(S_i^{mg,n}), \ \forall \boldsymbol{S}_i \in \mathcal{S}, \tag{5.21}$$

solves the optimization problem for user n in (5.16). Therefore, let $\psi^* \in \mathbf{\Psi}^{mg}$ be a max-gain scheduling adaptive transmission policy such that

$$\psi^*(\boldsymbol{S}_i, n) = \mu^{n,*}(\boldsymbol{S}_i^{mg,n}), \ \forall n \in \mathcal{N}, \forall \boldsymbol{S}_i \in \mathcal{S}. \tag{5.22}$$

Then policy ψ^* is a solution to the optimization problem in (5.15).

Remark: The idea behind the above discussion is quite simple and intuitive. When the scheduling rule, i.e., max-gain scheduling, does not depend on the buffer condition of any user, the transmission decisions applied to one particular user do not have any effect on the control of others. As a result, the problem of jointly controlling N users can be decoupled into N problems of controlling individual users.

5.5.3 Complexity of Obtaining and Implementing Max-gain Scheduling Optimal Transmission Policies

We note that in order to obtain a max-gain scheduling optimal transmission policy which is a solution to (5.15), we need to solve N reduced MDPs in (5.20). The number of system states in each of these reduced MDPs is $(B+1)K^N$. In general, this is simpler than solving an MDP of size $(B+1)^N K^N$ for the optimal adaptive scheduling/transmission policies discussed.

At the base station, implementing a max-gain scheduling optimal transmission policy is also simpler than doing so for an optimal adaptive scheduling/transmission policy. At the beginning of each time slot, instead of asking all N users to report their buffer occupancies, the base station first estimates the channel conditions of all users and decides which one will be allowed to transmit. Then, only this user will have to report the buffer condition to the base station so that his transmit power and rate can be determined.

Max-gain scheduling optimal transmission polices can also be implemented at each individual node. To do so, the base station broadcast the channel states of all N users at the beginning of each time slot. Then, the user with the best channel condition will decide what transmission rate and power to take, based on his buffer occupancy and the channel states of all users.

5.6 Round-robin Scheduling Optimal Transmission

5.6.1 Round-robin Scheduling Optimal Transmission Policies

In this section, let us consider a class of adaptive scheduling/transmission policies in which all N users are scheduled in a static round-robin manner. By static round-robin, we mean that the users are scheduled according to some fixed sequence, regardless of their buffer and channel conditions. If the advantage of max-gain scheduling is power efficiency, round-robin offers short-term fairness to the users. In particular, it satisfies the need to frequently transmit data of all users. Without loss of generality, we assume that user n is assigned time slots $iN + n - 1,\ n \in \mathcal{N},\ i \in \mathbb{N}$.

Definition 5.6.1. *A **round-robin scheduling adaptive transmission policy** is a policy $\psi = \{\phi_0, \phi_1, \ldots\},\ \psi \in \mathbf{\Psi}$, such that $\forall i \in \mathbb{N},\ \phi_i(\boldsymbol{S}_i, n) > 0$ only if $\mod(i, N) = n - 1$.*

Now let $\mathbf{\Psi}^{rr}$ denote the set of all round-robin scheduling adaptive transmission policies, we would like to solve the following problem:

$$\min_{\psi \in \mathbf{\Psi}^{rr}} \limsup_{T \to \infty} \frac{1}{T} \sum_{i=0}^{T} \sum_{n=1}^{N} \Big(\beta P(\phi_i(S_i, n), G_i^n) + L_o(B_i^n, \phi_i(S_i, n)) \Big). \tag{5.23}$$

5.6.2 Obtaining Round-robin Scheduling Optimal Transmission Policies

Similar to the case of max-gain scheduling, the problem in (5.23) can be decoupled into N optimization problems with the objective of the n^{th} problem is to find a round-robin scheduling adaptive transmission policy that minimizes the weighted sum of the packet loss rate and average transmit power for user n, i.e.,

$$\min_{\psi \in \mathbf{\Psi}^{rr}} \limsup_{T \to \infty} \frac{1}{T} \sum_{i=0}^{T} \Big(\beta P(\phi_i(S_i, n), G_i^n) + L_o(B_i^n, \phi_i(S_i, n)) \Big). \tag{5.24}$$

The optimization problem (5.24) can be simplified by the following observation. The operation of user n can be divided into frames, each contains N consecutive time slots. Frame i denotes the period from the beginning of time slot $Ni + n$ to the end of time slot $N(i + 1) + n - 1$. In the first time slot of each frame, user n is assigned the channel and can transmit at a positive rate. Then, in the following $N - 1$ consecutive time slots, user n is not scheduled and his transmission rate is set to zero. Therefore, in each frame, what really matter for the adaptive transmission of user n are his buffer occupancy at the beginning of the first time slot and the channel state during the first time slot. This motivates us to define the *effective* channel state, buffer occupancy, and transmission rate of user n in time frame i as

$$B_i^{rr,n} = B_{Ni+n}^n, \; G_i^{rr,n} = G_{Ni+n}^n, \text{and } U_i^{rr,n} = U_{Ni+n}^n. \tag{5.25}$$

The important thing to note is that, as $\{G_i^n\}$ and $\{B_i^n\}$ are Markov processes, so are $\{G_i^{rr,n}\}$ and $\{B_i^{rr,n}\}$. It is also straightforward to write down the dynamics of the effective buffer and channel states. In particular, the dynamics of $B_i^{rr,n}$ is

$$B_{i+1}^{rr,n} = \min\{B_i^{rr,n} - U_i^{rr,n} + A_i^{rr,n}, B\} \tag{5.26}$$

where

$$A_i^{rr,n} = \sum_{j=0}^{N-1} A_{Ni+n+j}^n \tag{5.27}$$

is the total number of packets arriving to the buffer of user n during time frame i. Let $\boldsymbol{V}$ be the set of all $(N+1)$-element vectors $\boldsymbol{v}$ such that $\boldsymbol{v}(0) = k$, $\boldsymbol{v}(N) = l$, $0 \leq k, l < K$, and $\boldsymbol{v}(1), \boldsymbol{v}(2), \ldots \boldsymbol{v}(N-1) \in \{0, 1, \ldots K-1\}$. The transition probabilities of the effective channel state of user 1 after each time frame can be written as

$$\begin{aligned} P_G^{rr}(k,l) = \Pr\{G_{i+1}^{rr,n} = l \mid G_i^{rr,n} = k\} &= \sum_{\boldsymbol{v} \in \boldsymbol{V}} \prod_{t=0}^{N-1} P_G(\boldsymbol{v}(t), \boldsymbol{v}(t+1)) \\ &= \sum_{\boldsymbol{v} \in \boldsymbol{V}} \Big(P_G(k, \boldsymbol{v}(1)) P_G(\boldsymbol{v}(N-1), l) \prod_{t=1}^{N-2} P_G(\boldsymbol{v}(t), \boldsymbol{v}(t+1)) \Big). \end{aligned} \tag{5.28}$$

Let us define the effective system state for user n at time i as

$$\boldsymbol{S}_i^{rr,n} = (B_i^{rr,n}, G_i^{rr,n}). \tag{5.29}$$

In addition, let $\mathbf{\Pi}_n^{rr}$ be the set of all transmission policies u that set the transmission rate for user n, i.e., $U_i^{rr,n} = \mu(S_i^{rr,n})$. Now, let $u^{n,*}$ be a (stationary) policy such that

$$u^{n,*} \;=\; \arg\min_{u\in\mathbf{\Pi}_n^{rr}} \limsup_{T\to\infty} \frac{1}{T}\sum_{i=0}^{T}\Big(\beta P(\mu(\boldsymbol{S}_i^{rr,n}), G_i^n) + L_o^{rr}(B_i^{rr,n}, \mu(\boldsymbol{S}_i^{rr,n}))\Big), \tag{5.30}$$

where $L_o^{rr}(B_i^{rr,n}, U_i^{rr,n})$ is the expected number of packets lost for user n during frame i, i.e.,

$$L_o^{rr}(B_i^{rr,n}, U_i^{rr,n}) = \mathbb{E}\Big\{\max\{0,\ B_i^{rr,n} - U_i^{rr,n} + A_i^{rr,n} - B\}\Big\}. \tag{5.31}$$

The expectation in (5.31) is with respect to $A_i^{rr,n}$ (defined in (5.27)).

Now let $\psi^* \in \mathbf{\Psi}^{rr}$ be a round-robin scheduling adaptive transmission policy that satisfies

$$\psi^*(\boldsymbol{S}_{iN+n-1}, n) = u^{n,*}(\boldsymbol{S}_i^{rr,n}),\ \forall n \in \mathcal{N},\ \forall i \in \mathbb{N}. \tag{5.32}$$

It can be shown that ψ^* is a solution to the optimization problem in (5.23).

5.6.3 Complexity of Obtaining and Implementing Round-robin Scheduling Optimal Transmission Policies

We note that the complexity of finding a round-robin scheduling optimal transmission policy is much smaller than those of the problems of finding optimal adaptive scheduling/transmission (Section 5.2.3), or max-gain scheduling optimal transmission policies (Section 5.5). In particular, the size of the MDP for each of the N users is $(B + 1)K$, which is the same as that of the MDP formulated for the single-user buffer/channel adaptive transmission considered in Chapter 3. It is also simple to implementing a round-robin scheduling optimal

Table 5.1: Channel states and transition probabilities.

Channel state k	0	1	2
γ_k	0	0.1	0.9
$P_G(k,k)$	0.6	0.6	0.6
$P_G(k,k+1)$	0.4	0.2	n.a.
$P_G(k,k-1)$	n.a.	0.2	0.4
$p_G(k)$	1/3	1/3	1/3

transmission policy. Especially, in order to determine the transmission rate of each user, no knowledge of other user channel or buffer condition is required, the control can be carried out by each individual user, instead of a centralized approach at the base station.

5.7 Numerical Results and Discussion

In this section, we numerically study the performance of different classes of buffer and channel adaptive scheduling/transmission policies considered in Sections 5.3, 5.4, 5.5, and 5.6. For convenient of notation, let us respectively denote by *Opt, MP, MG*, and *RR* the classes of optimal adaptive scheduling/transmission policies, max-product scheduling policies, max-gain scheduling optimal transmission policies, and round-robin scheduling optimal transmission policies.

5.7.1 System Parameters

The system parameters for our numerical study are as follows. The number of users in the system is set to $N = 2$. Data packets arrive to each user buffer according to a Poisson distribution with the average rate of $\lambda = 0.5$ packets per time slot. The buffer length of all users is set to 8 and 12 packets. The channel models are i.i.d. across users and are represented by a 3-state FSMC as in Tab. 5.1. As will be discussed, we also vary some parameters of the channel model to study their effects on the performance of different adaptive policies.

We assume that the power needed for a user to transmit reliably at rate u packets per time slot when his channel gain is γ_k is

$$P_w(u, k) = \frac{-WN_o(2^u - 1)}{1.5\gamma_k} \log(cP_b), \tag{5.33}$$

where $P_b = 10^{-3}$ is the required BER, $c = 0.5$ when $u = 1$ and $c = 5$ otherwise, $W = 100$ kHz is the channel bandwidth, and the power density of AWGN noise is $N_o/2 = 10^{-5}$ Watt/Hz. Note that $P_w(u, k)$ is the power needed for an uncoded M-ary quadrature amplitude modulation (MQAM) system with constellation size 2^u to have the BER of P_b when the channel gain is γ_k [GC97].

5.7.2 Performance of Different Adaptive Scheduling/ Transmission Schemes

For each class of adaptive scheduling/transmission policies, i.e., Opt, MP, MG, and RR, the performance metric we are interested in is the average packet loss rate (due to buffer overflow) versus the average transmit power. Here, the average packet loss rate is summed over all users and normalized by the total packet arrival rate. Note that the normalized system throughput is equal to one minus the normalized packet loss rate. The average transmit power is calculated per user and in each of the policies considered, all users consume the same average transmit power.

In Figs. 5.2, 5.3, 5.4, 5.5, we plot the performance, in terms of normalized packet loss rate versus average transmit power, of Opt, MP, MG, and RR for

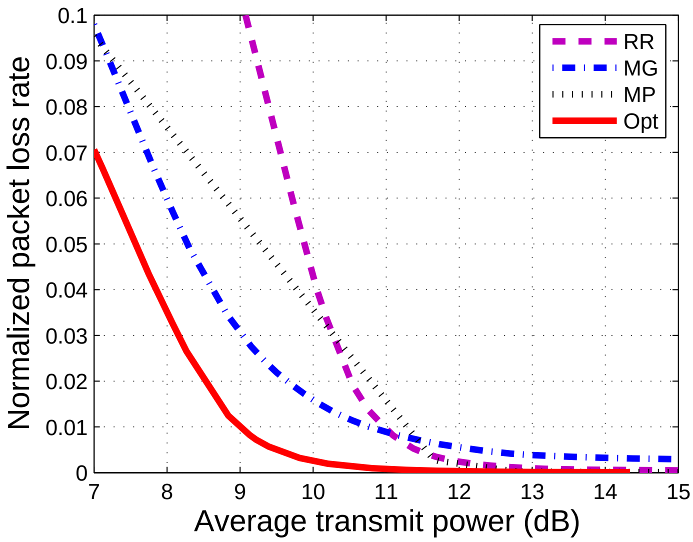

Figure 5.2: Performance, in terms of the normalized packet loss rate versus the average transmit power for different adaptive scheduling/transmission policies: Opt, MP, MG, RR. Number of users $N = 2$, data packets arrive at rate $\lambda = 0.5$ packets/time_slot with Poisson distribution, buffer length $B = 12$ packets, channel model is described in Tab. 5.1.

$N = 2$ and different values for other system parameters. This allows us to observe the general trends in the performance of the proposed classes of adaptive scheduling/transmission policies.

The first observation is that Opt always performs best. This is expected as in this class of policies, the scheduling and transmission decisions are jointly optimized. In general, the performance of MP varies considerably across the power range. At the high power range, the performance of MP is relatively close to that of Opt. However, at mid-range of average transmit power, MP performs quite far from optimal. MP does not offer a stable performance due to the inflexibility of the transmission scheme.

The performance of MG and RR follows opposite trends. At low power, MG outperforms RR and gets closer to the performance of Opt. On the other

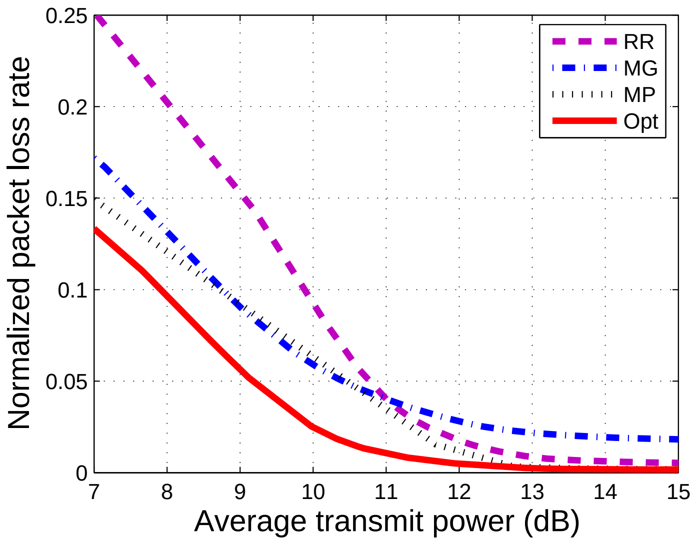

Figure 5.3: Performance, in terms of the normalized packet loss rate versus the average transmit power for different adaptive scheduling/transmission policies: Opt, MP, MG, RR. Number of users $N = 2$, data packets arrive at rate $\lambda = 0.5$ packets/time_slot with Poisson distribution, buffer length $B = 8$ packets, channel model is described in Tab. 5.1.

hand, in the high power range, RR performs much better than MG and closely approaches the performance of Opt. These trends in performance can be explained as follows. Max-gain scheduling is good for achieving power efficiency, as it allows a user with the best channel condition to transmit. As a result, the max-gain scheduling policies perform well at low range of transmit power, where the need for power efficiency is high. However, by favoring users with the best channel condition, max-gain scheduling is (short term) unfair to others. When transmission can be carried out at high power level, instead of power efficiency, what important is every user is allowed to transmit frequently. This is what RR does. Therefore, RR approaches optimal performance when average transmit power is increased.

To see how the relative performance of Opt, MG, and RR depend on dif-

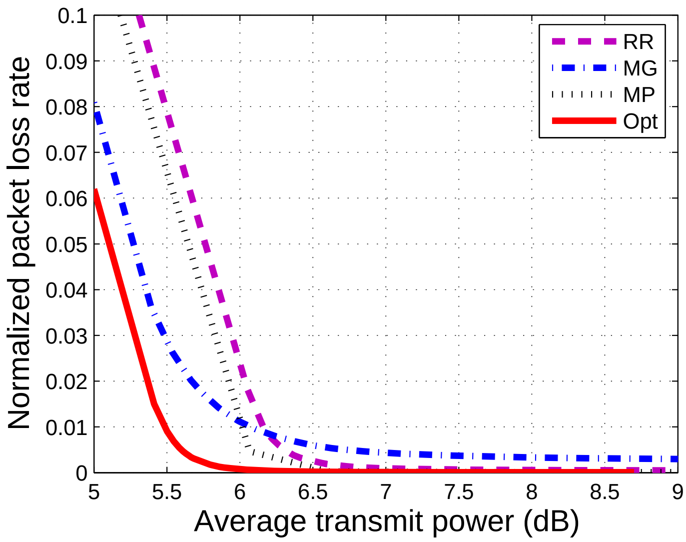

Figure 5.4: Performance, in terms of the normalized packet loss rate versus the average transmit power for different adaptive scheduling/transmission policies: Opt, MP, MG, RR. Number of users $N = 2$, data packets arrive at rate $\lambda = 0.5$ packets/time_slot with Poisson distribution, buffer length $B = 12$ packets, channel model is the same as in Tab. 5.1 except that the gains for states $\gamma_0, \gamma_1, \gamma_2$ are set to $0, 0.5, 0.9$ respectively.

ferent system scenario, we vary the system parameters and again compare the performance of these schemes in Fig. 5.3, 5.4, 5.5. First, in Fig. 5.3, we reduce the buffer size from 12 packets to 8 packets. As can be seen, all schemes perform worse. However, it seems that MG suffers more from the reduction in buffer length than RR does. This is shown by the fact that, when the buffer size is reduced, at low power, the gap between MG and RR is less while for the high power range, the gap between MG and RR is widen. This change in the relative performance is explained by the fact that, when there are less storage space, users are less capable of holding back data to wait for a good channel model, i.e., it is more costly to do max-gain scheduling. This also means that there is more advantage in scheduling every user regularly like in RR.

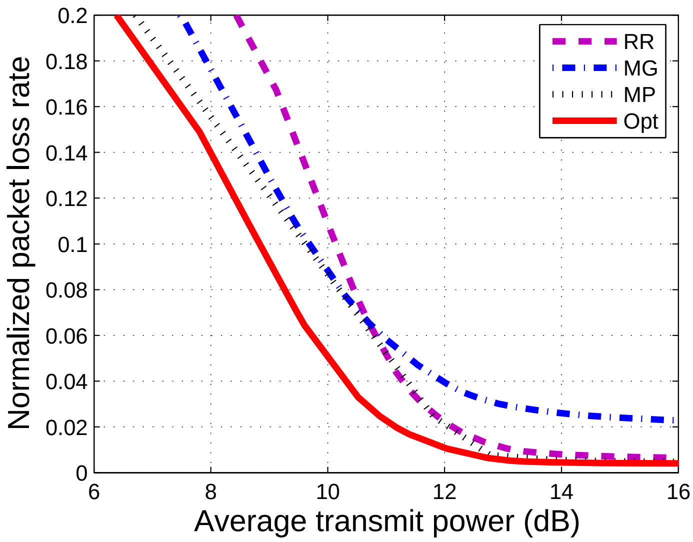

Figure 5.5: Performance, in terms of the normalized packet loss rate versus the average transmit power for different adaptive scheduling/transmission policies: Opt, MP, MG, RR. Number of users $N = 2$, data packets arrive at rate $\lambda = 0.5$ packets/time_slot with Poisson distribution, buffer length $B = 12$ packets, channel model is the same as in Tab. 5.1 except that the probability of staying in each channel state after each time slot is set to $P_G(k, k) = 0.8,\ k = 0, 1, 2$, probabilities of going up or down one channel state are equal.

Next, we look at the effect of changing the degree of fluctuation in the channel gains. In particular, if in Fig. 5.2, the channel gains in three states are 0, i.e. outage, 0.1, and 0.9, then in 5.4, we set these gain to 0, 0.5, 0.9. This means that there is less difference in the channel conditions in states 1 and 2. Now, there is less channel diversity for Opt and MG to take advantage of. On the other hand, RR, which schedules user without taking the channel conditions into account, will suffer less performance loss. This is shown in Fig. 5.4. As can be seen, the gap between MG and RR is less at the low power range and is more at the high power range.

The last parameter of concern is how fast the channel changes. We vary the frequency at which channel changes by changing the probability that the channel stay in each state after each time slot. In 5.2, this probability is set to 0.6 while it is increased to 0.8 in Fig. 5.5. This means the channel changes less frequently. When the channel changes slowly, max-gain scheduling will suffer, as some user with the best channel condition will hold the channel for long time, leaving others no chance to empty their buffers. As a result, the performance of MG decreases, relative to that of RR.

5.8 Hybrid Scheduling Schemes

5.8.1 Combined Round-robin and Max-gain Scheduling

From the system point of view, while round-robin scheduling offers fairness among users, the main objective of max-gain scheduling scheme is transmit power efficiency. Depending on the situation, it may be desirable to employ scheduling policies that offer a good balance between fairness and efficiency. This motivates us to look at the following hybrid scheduling scheme which is a combination of round-robin and max-gain scheduling. In particular, the N users in the system are divided into M separate groups. Let $N = N_1 + N_2 + \dots N_M$, where $N_1, N_2, \dots N_M$ are M positive integers, user n belongs to group m if and only if $\sum_{i=1}^{m-1} N_i < n \leq \sum_{i=1}^{m} N_i \quad 1 \leq n \leq N,\ 1 \leq m \leq M$. The Hybrid scheduling scheme selects a user to access the common channel in each time slot in two steps:

- *Step 1*: Select one user group among M user groups of the system in a round-robin manner.
- *Step 2*: Within the user group being selected in Step 1, select the user

with the best channel condition to access the common channel.

We term this hybrid scheme Hb_RR_MG. It is clear that the Hb_RR_MG scheme gives us a large degree of freedom to balance fairness and efficiency. As user groups are selected in a round-robin manner, no user is allowed to access the channel in two or more consecutive time slots. At the same time, within each group of users, the user who has the best channel condition is selected to access the channel, hence resulting in transmit power efficiency. It is also easy to see that, the hybrid scheduling is equal to round-robin scheduling when $M = N$ and max-gain scheduling when $M = 1$.

5.8.2 Combined Round-robin and Optimal Scheduling

Another hybrid scheduling scheme can be formed by combining round-robin scheduling and optimal scheduling. Similar to Hb_RR_MG, N users are divided into M groups and the groups are scheduled in a round-robin manner. However, when a user group is selected for a particular time slot, we optimally schedule one of the users, based on the channel and buffer conditions of other users in the group. We term this scheme Hb_RR_Opt. Note that Hb_RR_Opt has higher complexity than Hb_RR_MG, but also offers better performance.

5.8.3 Hybrid Scheduling Optimal Transmission Policies

Conditioned on a hybrid scheduling scheme, the optimal adaptive transmission scheme can be derived in a straightforward manner, by applying the approaches in Sections 5.3, 5.5, and 5.6. In particular, like the case of the round-robin scheduling, each user is controlled in a frame-by-frame basis, with each frame consisting of M consecutive time slots. Using the same approach as in Section

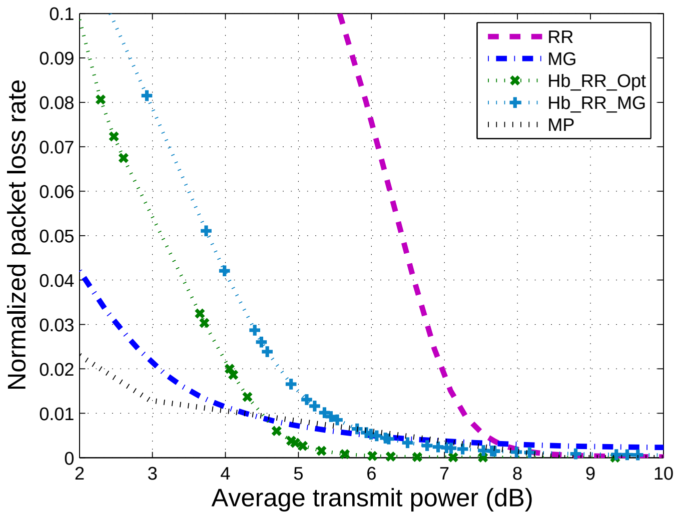

Figure 5.6: Performance, in terms of the normalized packet loss rate versus the average transmit power for different adaptive policies: MG, RR, MP, Hb_RR_Opt, and Hb_RR_MG. Number of users $N = 4$, data packets arrive at rate $\lambda = 0.25$ packets/time_slot with Poisson distribution, buffer length $B = 12$ packets, channel model is described in Tab. 5.1.

5.6, with a frame length of M instead of N time slots, we derive the channel transition probabilities for each user after each time frame. Now, using the same approach as in Sections 5.3, 5.5, we can derive the optimal adaptive transmission policy for a user in group m when either Hb_RR_Opt or Hb_RR_MG is employed.

5.8.4 Performance of Hybrid Scheduling Optimal Transmission Policies

Let us look at the performance of the Hb_RR_MG scheme first. Note that this scheme is a combination of MG and RR. As MG and RR are good at either low or high power regions, but not both, the aim of HB is to bridge the gap. In Figs. 5.6 and 5.7, we plot the performance of Hb_RR_MG, MG, and RR,

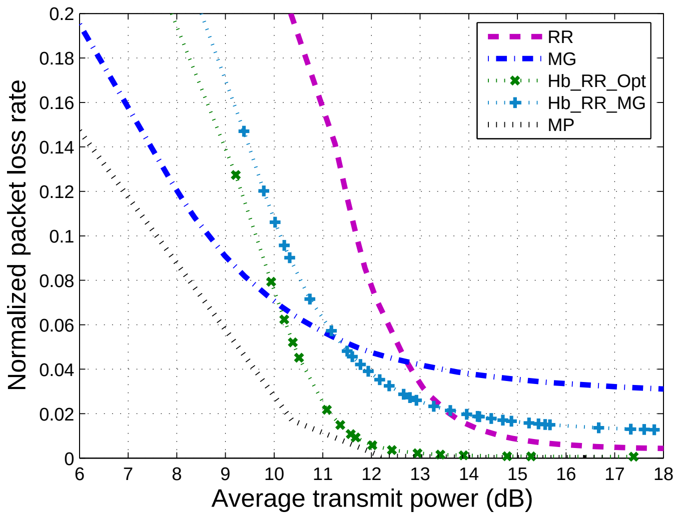

Figure 5.7: Performance, in terms of the normalized packet loss rate versus the average transmit power for different adaptive policies: MG, RR, MP, Hb_RR_Opt, and Hb_RR_MG. Number of users $N = 4$, data packets arrive at rate $\lambda = 0.5$ packets/time_slot with Poisson distribution, buffer length $B = 12$ packets, channel model is described in Tab. 5.1.

in terms of normalized packet loss due to buffer overflow versus the average transmit power. Here, there are $N = 4$ users in the system and for Hb_RR_MG the 4 users are divided into $M = 2$ groups, each consists of 2 users. The packet arrival rate is set to $\lambda = 0.25$ and 0.5 packets per time slot. As can be seen, Hb_RR_MG offers a good balance between MG and RR. In particular, at mid-range power levels, HB outperforms both MG and RR.

In Figs. 5.6 and 5.7, we also show the performance of the Hb_RR_Opt scheme. Note that this scheme is the combination of Opt and RR. As can be seen, Hb_RR_Opt policies always outperform RR policies. When compared to the performance of MG, Hb_RR_Opt offers better performance in the mid range and high range of transmit power.

5.9 Observations and Conclusions

After studying the performance and complexities of different adaptive scheduling and transmission policies for our multiple-access system, we make the following observations:

- Firstly, if the scheduling decisions depend on the channel and buffer conditions of all users, then the optimal transmission policy of each user must also take the buffer and channel conditions of all other users into account. This results in a relatively high computational complexity as well as signaling cost to obtain and implement optimal adaptive scheduling/transmission policies.

- For those scheduling policies that only take the channel conditions into account, max-gain scheduling is good for the range of low of power constraint, and when there are long buffers. When plenty of transmit power is available, or when there is limited buffer space, round-robin scheduling is a better choice. Note that max-gain, round-robin, and hybrid schemes do not require knowledge of all buffer occupancies. This can greatly reduce the amount of signaling required to implement adaptive scheduling/transmission.

- When the channel and data arrival statistics are not available, max-product scheduling is a reasonable choice. Note that with max-product scheduling, the complexity to obtain and implement optimal adaptive transmission is as high as that of obtaining and implementing optimal joint scheduling/transmission. Therefore, it is reasonable to not consider adaptive power control when max-product scheduling is implemented.

Cross-layer design, while promising significant performance improvement for energy constrained wireless communication systems, can come with high complexity in design and analysis. This general statement is also true for our problem of cross-layer adaptive scheduling/transmission to maximize the system throughput, subject to average power constraints. Therefore, instead of focusing on optimal adaptive scheduling/transmission policies, the main objective of this chapter is to look at the trends in performance of different classes of suboptimal policies. The important contribution of this chapter is to identify how different classes of suboptimal policies perform, in comparison to the optimal performance.

CHAPTER 6

JOINT SCHEDULING, TRANSMISSION, AND SOURCE COMPRESSION IN SENSOR NETWORKS

The problems of buffer and channel adaptive transmission and scheduling studied in Chapters 3, 4, and 5 focus heavily on adapting to different sources of variations in the parameters of the MAC and PHY layers. In this chapter, we will demonstrate that cross-layer design is still highly beneficial at the MAC and PHY layers, even when there are no variations and randomness in the system parameters.

We propose a novel approach that exploits the broadcast nature of the wireless medium for energy conservation in spatially correlated wireless sensor networks. Since wireless transmission is inherently broadcast, when one sensor node transmits, other nodes in its coverage area can receive the transmitted data. When data collected by different sensors are correlated, each sensor can utilize the data it overhears from other sensors to compress its own data and conserve energy in its own transmissions.

We apply this idea to a class of cluster-based wireless sensor networks in which each sensing node transmits collected data directly to its cluster head using time division multiple access (TDMA). We formulate the problem in which sensors in each cluster collaborate their transmitting, receiving, and compressing activities to optimize their lifetimes. From the system design point of view, the problem considered in this chapter deals with scheduling, transmission, reception, and data compression in an integrated manner. Therefore, it can be characterized as an instance of cross-layer design.

The main results of this chapter can be summarized as follows.

- We propose the collaborative broadcasting and compression (CBC) approach which allows sensors to cooperate in transmitting, receiving, and compressing data in order to conserve energy (Section 6.4).
- We formulate an optimization problem of which the objective is to find a CBC scheme that jointly optimizes the lifetimes of all sensors in each cluster, with respect to some optimality criteria (Section 6.5). We show that this lifetime optimization problem can be solved by a sequence of linear programming problems (Section 6.6).
- When the number of sensors in each cluster is large, we propose a heuristic CBC scheme that achieves near optimal performance at a lower complexity (Section 6.7).
- We discuss how our CBC schemes perform under nodes' startup cost and transmission errors and also argue that the schemes are nearly independent to the operation of the relaying network (Section 6.8).
- Finally, we obtain numerical results which show that by applying the CBC approach, significant increase in sensor lifetime can be achieved (Section 6.9).

We note that some of the above results have been presented in [HM05a, HM05e, HM05b].

6.1 Motivations

Recent advances in wireless communication and microelectronics have enabled the possibility of wireless sensor networks (WSNs) [ASSC02], which can consist of hundreds or even thousands of low-cost, low-power, and small-in-size sensors. As these cheap and tiny sensors can only be equipped with small batteries, and in many applications, battery recharging/replacing is not desirable, achieving energy-efficiency to increase sensors' lifetimes is an important design criterion for WSNs.

In many sensing networks, a high degree of spatial correlation exists among the readings of different sensors. By allowing nodes to cooperate to carry out joint data compression and aggregation, the amount of data communicated within the network can be reduced. This can help conserve energy and extend sensors' lifetimes.

The work in this chapter deals with removing the redundancy due to spatial correlation among nodes in WSNs. The novelty of our work lies in the fact that we exploit the inherent broadcast nature of the wireless medium for nodes to share and jointly compress their data. The core idea is that, when one node broadcasts data, other nodes within the transmission range can receive and utilize this data in compressing their own data.

We note that most of the works concerning data compression and aggregation in WSNs adopt a common model for the wireless medium, in which a wireless channel is abstracted as a single point-to-point link between a pair of nodes, e.g., [PR99, IGE00, SS02a, KEW02, GE03, CPR03, CLV04]. This point-to-point link model, while simplifying design and analysis, ignores important advantages that come with the inherent broadcast nature of the wireless medium. We contend that the wireless broadcast property offers nodes in a WSN much more freedom in carrying out joint data compression and achieving energy efficiency.

To illustrate our point, let us consider a simple system of four wireless sensor

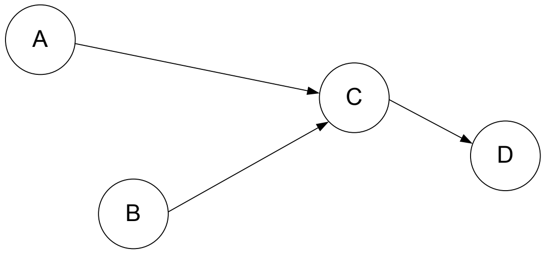

Figure 6.1: System of four wireless sensor nodes. Each wireless channel is abstracted as a single point-to-point link. Transmission from (A) to (C) does not reach (B). (A) and (B) can only carry out joint data compression by following the complex distributed source coding approach.

nodes (A), (B), (C), and (D) depicted in Figs. 6.1 and 6.2. Nodes (A) and (B) need to transmit collected data to (C), who then relays the data toward (D). Note that both Figs. 6.1 and 6.2 represent the same network. The only difference is that in Fig. 6.1, the wireless broadcast property is not considered, while this is taken into account in Fig. 6.2.

In Fig. 6.1, when either (A) or (B) transmits to (C), the other node does not receive and decode the transmitted data. With this point-to-point link model, the only way for (A) and (B) to jointly compress their data is to carry out distributed source coding [SW73, WZ76, PR99]. In Fig. 6.2, we suppose that all nodes transmit using omni-directional antennas under the free-space path loss model. Let (A) transmit its data to (C) first before (B) does. Furthermore, assuming that the distance between (A) and (B) is not more than that between (A) and (C), then when (A) transmits to (C), its data can be received by (B). Now, if (B) receives the data of (A), it can utilize these data in carrying out data compression. More specifically, node (B) can compress its data based on the explicit knowledge of the data of (A), and therefore, avoid the complexity

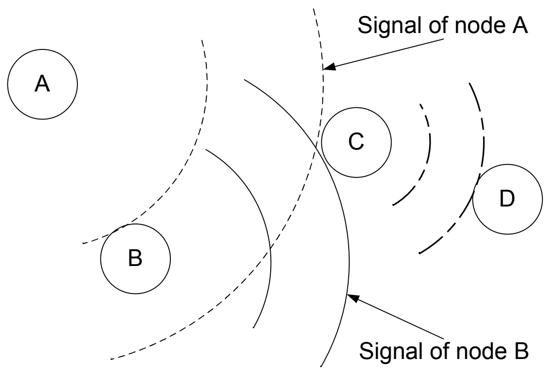

Figure 6.2: All nodes transmit using omni-directional antennas. Assuming the distance between (A) and (C) is not less than the distance between (A) and (B), then (B) can capture data sent from (A) to (C) and then uses that data to compress its own data.

associated with implementing distributed source coding.

The above observations motivate us to exploit the wireless broadcast property for nodes to carry out joint data compression in a spatially-correlated cluster-based wireless sensor network. In our network model, there are two types of nodes, i.e., sensing nodes and cluster head/relaying nodes. Each cluster consists of multiple sensing nodes (also called sensors) and one cluster head. Data collected by each sensor are forwarded to the corresponding cluster head using direct transmission and time division multiple access (TDMA). The cluster head in turn routes data collected in its clusters toward a command center which can be accessed by the end users. This network is depicted in Fig. 6.3. We also assume that sensing nodes are much more energy constrained compared to cluster head/relaying nodes. The objective is therefore to conserve energy and prolong lifetimes of sensing nodes in each cluster. This is achieved by scheduling the data transmission and reception for sensor nodes in each cluster so that they can carry out joint data compression in an efficient manner.

6.2 Related Work

Works that are most closely related to the problem considered in this chapter are by Chou et al. [CPR03], Agnihotri et al. [ANJ05], and by Scaglione and Servetto [SS02a]. In [CPR03], the authors propose an approach that combines adaptive signal processing and distributed source coding for sensor nodes in cluster-based WSNs to conserve energy. The main idea of [CPR03] is to let sensors in each cluster blindly compress their data with respect to one another, but without the need of explicit inter-sensor communication. The processing burden in this case is shifted to the cluster heads, who need to perform decoding with side information and adaptive filtering to estimate relevant correlation. In [ANJ05], the authors follow a similar approach of implementing distributed source coding for sensors to conserve transmission energy. Their objective is also similar to ours, i.e., to maximize the lifetime of the sensor who dies first. For more details on distributed source coding, please refer to [SW73, WZ76, PR99]. In [SS02a], the authors propose an approach which is opposite to that of [CPR03] and [ANJ05]. In particular, they promote the idea of source coding based on explicit data of other nodes in the network, which are made available through routing. They argue that, as a routing scheme is already implemented in a WSN, nodes in each routing path actually have explicit information of some other nodes, and therefore, they can carry out classical source coding and avoid the complexity of distributed source coding.

The approach proposed in this chapter combines the advantages of both [CPR03], [ANJ05], and [SS02a] while avoiding their disadvantages. On one hand, like [CPR03] and [ANJ05], we allow nodes in each cluster to carry out data compression with respect to one another, and without any extra inter-sensor transmissions. The core idea here is the realization that as one node transmits its data to the cluster head, due to the broadcast nature of the media, its transmission reaches multiple other nodes. In addition, as nodes carry out compression based on the explicit information that they receive when other nodes broadcast, classical source coding can be employed as in [SS02a].

It is useful to further elaborate on the differences between our data compression approach and the distributed source coding approach in [SW73, WZ76]. In terms of the enhanced network lifetime, the performance of our approach is upper-bounded by the performance of the distributed source coding approach. This is due to two reasons. Firstly, we constrain that each sensor can only compress based on the data of at most one other sensor (see Section 6.5.2). This makes it not possible for CBC to achieve the optimal joint-entropy coding rate of distributed source coding. Secondly, in our CBC scheme, before compressing, each sensor needs to spend some extra energy to receive from another sensor, this receiving energy is not required for distributed source coding. In terms of implementation complexity, our approach is much easier to implement compared to the distributed source coding approach. In particular, as CBC allows sensors to compress their data based on explicit knowledge of other sensors' data, simple compression scheme such as differential encoding can be employed. On the other hand, to implement distributed source coding is highly complex and requires exact knowledge of the correlation statistics.

The idea of exploiting the broadcast nature of wireless media for wireless ad hoc networks has been proposed in [WNE00, SSZ01, DMS+03]. In [WNE00], Wieselthier, Nguyen, and Ephremides propose a broadcast incremental power (BIP) algorithm for minimum-power tree for wireless networks. This idea is then further developed in [SSZ01, DMS+03]. In this chapter, we apply this philosophy of exploiting the wireless broadcast advantage to achieve energy savings for wireless sensor networks.

6.3 Model of A Cluster-based Wireless Sensor Network

6.3.1 Network Architecture

We consider a small-to-medium-sized cluster-based wireless sensor network as shown in Fig. 6.3. Sensor nodes are organized into clusters and each cluster is responsible for monitoring a geographical area. We adopt a heterogeneous model in which there are two types of nodes. Type I nodes are sensors whose responsibility is to sense the surrounding environment and then transmitting collected data directly to cluster heads who are type II nodes. Type II nodes gather/aggregate the data collected in their corresponding clusters and relay them toward a command center. We assume that type II nodes are less energy-constrained than type I nodes. We note that the algorithms presented in this paper will work with any clustering algorithms in which nodes are clustered based on having correlated data. For the numerical analysis in Section 6.9, we cluster nodes based on their location. In particular, each sensor will be associated with the closest cluster head. Note also that in Fig. 6.3, broadcast communication always takes place and the transmission of one node can be received by every node in the coverage area. The arrows are used to indicate intended destinations only. As will be explained in Section 6.3.4, our assumption of direct transmission toward cluster heads is suitable for WSNs with small to medium cluster size.

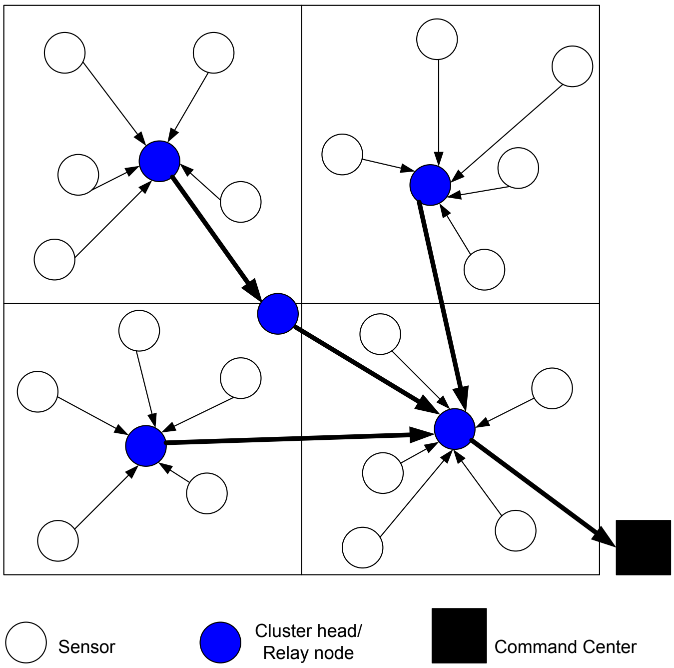

Figure 6.3: Model of a cluster-based wireless sensor network. There are two types of nodes, i.e. sensing nodes (type I) and data-gathering/ relaying nodes (type II). Sensing nodes transmit collected data directly to the corresponding cluster heads, who then route the data toward a command center.

6.3.2 Sensing and Communication

We consider a periodic sensing scenario in which time is divided into intervals of equal duration called *data-gathering rounds*. In each data-gathering round, each sensor collects useful information about the surrounding environment and outputs a data packet. The data are then forwarded toward the command center using the following mechanism.

- Within each cluster, sensors send data directly to the cluster head using time division multiple access (TDMA). In particular, the duration of each round is divided into slots and each sensor is assigned one slot to transmit data. We assume that inter-cluster interference is negligible. One way to achieve this is by assigning non-overlapping frequency bands to adjacent clusters.

- Upon receiving data collected in their clusters, cluster heads carry out the necessary data fusion/aggregation tasks. After that, the processed data is routed toward the command center over the relay network formed by all type II nodes.

We note that TDMA has been chosen in a number of WSN implementations [HCB00, SCI$^+$01] due to its simplicity, low overhead, short communication duty cycle, and no packet collisions. All these factors help conserve sensor nodes' energy. However, it should be noted that TDMA is only effective for scenarios in which the number of transmitting nodes is relatively stable over time. This is, in fact, true in our model of data-gathering WSN. For other sensing applications in which the number of active nodes change frequently, such as those event-based WSN, a contention-based approach would be more scalable than TDMA.

Before moving on, we would like to highlight the fact that, within each cluster, our system model is very similar to the multiple-access model considered in Chapters 2 and 5. The major difference is that in this chapter, we are focusing on exploiting the correlation among data collected by different sensor nodes.

6.3.3 Energy Model for Wireless Sensor Nodes

First of all, we assume that the sensing operation of each sensor consumes a fixed amount of energy during each data-gathering round. In order to achieve energy-efficiency for sensors, we only focus on controlling their communication-related activities. For the communication-related energy consumption, we adopt the first-order energy model used in [HCB00, HCB02]. In particular:

- The energy consumed to receive r bits is

$$E_{rx}(r) = E_e r \tag{6.1}$$

where E_e (in Joules/bit) is the energy consumed in the electronic circuits of the transceiver when receiving or transmitting one bit of information. Typical values for E_e range from 10nJ/bit to 100nJ/bit.

- The energy consumed to transmit r bits over a distance of d meters is

$$E_{tx}(r, d) = E_e r + E_a d^{\alpha} r \tag{6.2}$$

where α is the channel loss exponent which is typically in the range $2 \leq \alpha \leq 4$. For short communication distances, a free-space path loss model can be assumed and $\alpha = 2$. As the distance increases, a multipath model is more appropriate and $\alpha = 3$ or 4 [Rap96]. E_a (in Joules/bit/m$^{\alpha}$) is the energy consumed in the power amplifier to transmit one bit of information over a distance of one meter.

E_a depends on the receiver sensitivity and its range is from 10pJ/bit/m^2 to 100pJ/bit/m^2 for the free-space path loss model.

- The energy consumed to compress r bits is

$$E_{cp}(r) = E_c r. \tag{6.3}$$

where E_c (in Joules/bit) is the energy used by the processor to compress one bit of information in a data packet based on given side information. In general, E_c is much smaller than the electronic energy E_e. We note that a more complicated model for the compression energy could take into account various factors such as compression ratio and the amount of side information.

6.3.4 Direct Transmission versus Multihopping

At this point, let us justify our assumption of direct data transmission from sensors toward corresponding cluster heads. Note that the same assumption has been made in some related WSN works, i.e., [HCB00, ML02, CPR03, ANJ05]. In small-to-medium-sized WSNs (which is our assumption), due to short distance between nodes, the energy consumed for receiving is comparable to what is consumed for transmitting a given amount of data. In such scenarios, it has been pointed out in [HCB00] that direct transmission is in fact more energy-efficient than multihop routing. Let us demonstrate this fact based on a simple network in Fig. 6.4.

In Fig. 6.4, node (A) needs to communicate r bits to cluster head (C). If (A) transmits the data directly to (C), from Section 6.3.3, the total energy consumption would be:

$$E_{direct} = E_e r + E_a d_{AC}^2 r, \tag{6.4}$$

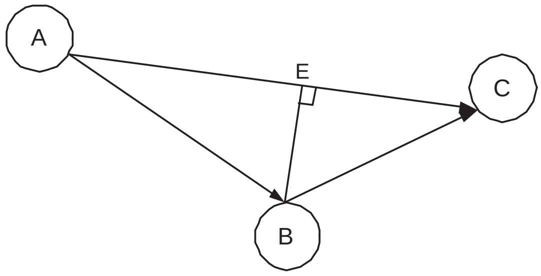

Figure 6.4: A simple network with two sensors (A) and (B) communicating to cluster head (C).

where d_{AC} is the distance (in meters) between (A) and (C).

Now, consider using node (B), which lies somewhere in between (A) and (C), to relay data from (A) to (C). In that case, the total energy consumed to transmit r bits from (A) to (B), and then from (B) to (C) would be:

$$\begin{aligned} E_{two\_hop} &= (E_e r + E_a d_{AB}^2 r) + (E_e r) + (E_e r + E_a d_{BC}^2 r) \\ &= 3E_e r + E_a (d_{AB}^2 + d_{BC}^2) r. \end{aligned} \tag{6.5}$$

Note that when (B) lies in between (A) and (C) as in Fig. 6.4, we have:

$$d_{AB}^2 + d_{BC}^2 \geq d_{AE}^2 + d_{EC}^2 = 0.5(d_{AC}^2 + (d_{AE} - d_{EC})^2) \geq 0.5 d_{AC}^2. \tag{6.6}$$

From (6.4), (6.5), (6.6), it can be seen that, for the network in Fig. 6.4, direct transmission will be more energy-efficient than two-hop routing when:

$$d_{AC} < 2\sqrt{E_e/E_a}. \tag{6.7}$$

As an example, if we select some typical values as $E_e = 50nJ/bit$, $E_a = 100pJ/bit/\text{m}^2$, then when $d_{AC} < 45m$, it is more energy-efficient to employ direct communication than two-hop routing. In other words, the assumption of direct transmission from sensors to cluster heads is reasonable in our model of small-to-medium-sized WSNs.

6.3.5 Spatial Correlation and Data Compression

The problem considered in this chapter aims to exploit the spatial correlation among sensor readings for nodes to carry out data compression. In that light, it is appropriate to discuss how spatial correlation and data compression are related.

First we discuss a statistical/information-theoretic approach for specifying spatial correlation and data compression. In this approach, the readings at each sensor are regarded as samples of a random variable and the correlation among readings at different sensors are characterized in an exact way, i.e., by specifying their joint probability distribution [DGM+04], or by establishing the relationship among the random variables [JP04], or by determining their joint entropy [PKG04]. Given a spatial correlation model, the conditional entropy of the quantized data of one sensor, given knowledge of some other sensors's data, can be computed. In general, it is expected that the conditional entropy will decrease when nodes get closer. Using entropy coding, sensors can then compress and transmit at a rate equal to the corresponding conditional entropy.

Now let us consider a more practical approach which is useful when all sensors measure continuous values in the same range and then employ the same quantization scheme. For sensors that are close to one another, the difference in their quantized measures can be small. In that case, simple differential encoding can be employed, i.e., when a node knows the quantized measure of another node, it will only transmit the difference with respect to that measure. This is suboptimal to the approach of characterizing the joint entropy and employing entropy encoding discussed above. However, it has the advantage of not requiring nodes to know the exact spatial correlation structure.

Using either entropy coding or differential encoding as described above, a sensor can compress its data based on the data of another node and therefore, eliminate or reduce the redundancy due to spatial correlation. This will allow the compressing node to transmit less data in a data-gathering round.

6.4 Collaborative Broadcasting and Compression: A Simple Case

6.4.1 A Simple Cluster-based Sensor Network

Let us introduce our approach by considering a simple cluster-based WSN depicted in Fig. 6.2. This network consists of only one cluster, which is composed of two sensors (A) and (B) and the cluster head (C), which gathers data collected by (A) and (B) and routes them toward the command center (D). We assume that all nodes transmit using omni-directional antennas and a free-space path loss scenario ($\alpha = 2$). By studying this simple network, we will illustrate the main concepts of our approach. A more general network will be considered in Sections 6.5, 6.6, and 6.7.

If the distance between (A) and (B) is not more than that between (A) and (C), then when (A) transmits to (C), its transmission can also be received by (B). Node (B) therefore has the option of first receiving the data of (A) and then using these data to compress its own data. If (B) does so, for the sake of brevity, we simply say (B) *compresses based on* (A). In addition, we refer to the approach in which sensor nodes coordinate their transmission and reception activities in carrying out joint data compression as *collaborative broadcasting and compression (CBC).*

6.4.2 Incentives for Collaboration

Let d_{AC}, d_{BC}, and d_{AB} denote the distances (in meters) between (A) - (C), (B) - (C), and (A) - (B) respectively. For this section, we assume that $d_{AB} \leq d_{AC}$. Let r_A and r_B be the amounts of uncompressed data (in bits) that (A) and (B) need to send to (C) during each data-gathering round. Furthermore, let $r_{B|A}$ be the amount of data that B needs to transmit to (C) if it compresses based on (A).

Using (6.1), (6.2), and (6.3), the energy (B) consumes to transmit r_B bits to (C) without compressing based on (A) is

$$E_B = E_e r_B + E_a d_{BC}^2 r_B. \tag{6.8}$$

On the other hand, the total energy that (B) will spend if it receives from (A), compresses based on (A), and finally transmits $r_{B|A}$ bits to (C) is

$$E_{B|A} = E_e r_A + E_c r_B + E_e r_{B|A} + E_a d_{BC}^2 r_{B|A}. \tag{6.9}$$

To make it easier to identify the incentives for (B) to compress based on (A), assume that $r_A = r_B = R$ while $r_{B|A} = r,\ r \leq R$. Then from (6.8) and (6.9), node (B) will save energy by compressing based on (A) when

$$\frac{r}{R} \;<\; \frac{E_a d_{BC}^2 - E_c}{E_a d_{BC}^2 + E_e}. \tag{6.10}$$

We call $\frac{r}{R}$ the *compression ratio* as it is the ratio of the compressed and uncompressed amounts of data that (B) sends to (C). Node (B) can choose its compression ratio based on a variety of factors, including requirements on acceptable distortion at the receiver. Based on (6.8), (6.9) and (6.10), we note that there is more incentive for (B) to compress based on (A) when

- d_{BC} increases, i.e., node (B) moves farther from the cluster head. In fact, it is evident from (6.10) that there is a value of d_{BC} below which compression is ineffective, i.e., node (B) will spend more energy to compress and transmit than not to compress at all.

- $\frac{r}{R}$ is small, i.e., a significant reduction in the size of the data of (B) can be achieved by compression.

- node (B) consumes less energy due to the transceiver electronics and the processor, i.e., when E_e and E_c decrease.

We illustrate the above observations by using the following numerical values: $E_a = 100$ pJ/bit/m^2, $E_c = 5$ nJ/bit, and $E_e = 10,\ 50,\ 100,\ 200$ nJ/bit. Fig. 6.5 shows the boundary of the region when it is beneficial for (B) to compress based on (A). Specifically, the area below each curve corresponds to the values of compression ratio $\frac{r}{R}$ and transmission distance d_{BC} at which (B) should compress based on (A).

6.4.3 Maximizing the Lifetime of the Node Who Dies First

In this section, we consider the problem of finding the control scheme that maximizes the time until one of the sensors in a cluster dies. For the network in Fig. 6.2, we have two possible CBC policies:

- *Policy* μ_1: Let (A) transmit to (C) first, (B) chooses either to transmit uncompressed data to (C) or, if it is beneficial, to compress based on (A) and then transmits to (C).

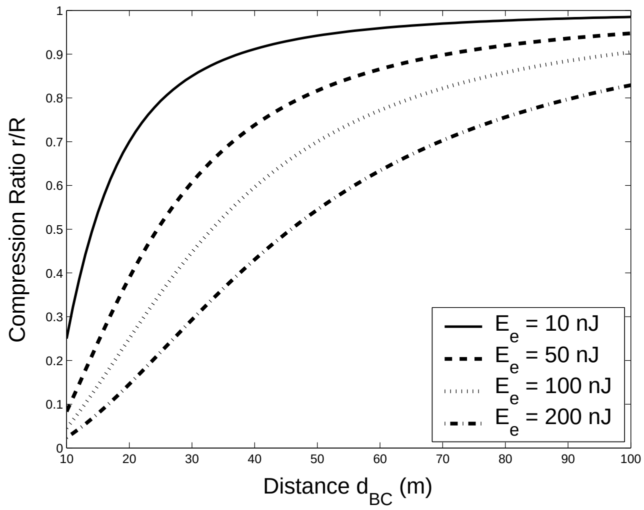

Figure 6.5: The incentives for node (B) to compress based on (A) (for the network in Fig. 6.2). $E_a = 100$pJ/bit/m^2, $E_c = 5$nJ/bit and $E_e = 10,\ 50,\ 100,\ 200$nJ/bit. The area below each curve corresponds to the region in which (B) can save energy by compressing based on (A).

- *Policy* μ_2: Let (B) transmit to (C) first, (A) chooses either to transmit uncompressed data to (C) or, if it is beneficial, to compress based on (B) and then transmits to (C).

For policy μ_1, the energy consumed by (A) will be

$$E_A^{\mu_1} = E_A = E_e r_A + E_a d_{AC}^2 r_A, \tag{6.11}$$

while the energy consumed by (B) will be

$$E_B^{\mu_1} = \min\left\{E_B, \frac{E_{B|A}}{\mathbb{1}(d_{AB} \leq d_{AC})}\right\}. \tag{6.12}$$

In (6.12), $\mathbb{1}(.)$ denotes the indicator function, which returns 1 if the expression inside the brackets is true and returns 0 otherwise. Note that (B) can compress based on (A) only when $d_{AB} \leq d_{AC}$, if $d_{AB} > d_{AC}$ then $\frac{E_{B|A}}{\mathbb{1}(d_{AB} \leq d_{AC})} = +\infty$ and (6.12) gives $E_B^{\mu_1} = E_B$.

Similarly, when policy μ_2 is applied, we can write the energy consumption of (A) and (B) as:

$$E_A^{\mu_2} = \min\left\{E_A, \frac{E_{A|B}}{\mathbb{1}(d_{AB} \leq d_{BC})}\right\} \tag{6.13}$$

$$E_B^{\mu_2} = E_B. \tag{6.14}$$

Note that in (6.13)

$$E_{A|B} = E_e r_B + E_c r_A + E_e r_{A|B} + E_a d_{AC}^2 r_{A|B} \tag{6.15}$$

where $r_{A|B}$ is the amount of data that (A) needs to transmit if it compresses based on (B).

Let e_A and e_B be the initial energies of (A) and (B) respectively. The problem of maximizing the time until at least one of the nodes (A) and (B) dies can be formulated as:

$$\arg\max_{t_1, t_2} \{t_1 + t_2\} \tag{6.16}$$

subject to:

$$t_1 \geq 0, \quad t_2 \geq 0, \tag{6.17}$$

$$E_A^{\mu_1} t_1 \ + \ E_A^{\mu_2} t_2 \ \leq \ e_A, \tag{6.18}$$

$$E_B^{\mu_1} t_1 \ + \ E_B^{\mu_2} t_2 \ \leq \ e_B. \tag{6.19}$$

Here t_1 and t_2 are the total numbers of data-gathering rounds that policies μ_1 and μ_2 are employed respectively. The constraints in (6.17) are obvious. The constraint in (6.18) basically means that the total energy consumed by sensor (A) during $t_1 + t_2$ data-gathering rounds cannot exceed its initial energy storage. Similar explanation is for constraint (6.19). We note that the order in which μ_1 and μ_2 are employed is not important. Also, for μ_1, the t_1 data-gathering rounds that this policy is employed do not have to be contiguous. The same is true for μ_2. In addition, t_1, t_2 must take integer values and the above optimization problem is an integer linear program. However, for applications in which sensors' lifetimes are much longer than each data-gathering round, t_1, t_2 can be treated as real variables. Then the above optimization is a linear programming problem and can be solved efficiently with standard methods [HL95].

6.5 Collaborative Broadcasting and Compression: A general network

We now apply the CBC approach to control a general cluster-based sensor network as depicted in Fig. 6.3. Note that our control will still be carried out within each cluster.

6.5.1 General Notation

First of all, we ask the reader to bear in mind that the notation used in this chapter is independent to that used in Chapters 2, 3, 4, and 5. We consider a cluster composed of K sensors and a cluster head. The sensor nodes are numbered from 1 to K and the cluster head is denoted by H. Let us introduce the following notation:

- $\mathbf{N} = \{1, \ldots K\}$ is the set of all sensors in the cluster.
- $d_{ik},\ i, k \in \mathbf{N}$, is the distance (in meters) between sensor i and sensor k. d_{kH} is the distance between sensor k and the cluster head.
- $\mathbf{N}_k = \{i \in \mathbf{N},\ i \neq k \mid d_{ik} \leq d_{iH}\},\ k \in \mathbf{N}$, is the set of all nodes whose transmission to the cluster head can be received by k.

- $e_k,\ k \in \mathbf{N}$, is the initial energy of sensor k.
- $E_k(i),\ k, i \in \mathbf{N},\ i \neq k$, is the total energy consumed by node k in each round when it compresses based on i. We also use $E_k(0)$ to denote the energy consumed by k when it does not compress based on any other node. Note that $E_k(i)$ can be determined using (6.1), (6.2), (6.3).

6.5.2 Control During Each Data-gathering Round

During each data-gathering round, in order to specify how nodes collaborate their data transmission and compression, two control decisions must be made. Firstly, a transmission order needs to be specified, i.e., each sensor should be assigned a time slot for data transmission. Secondly, given the transmission order, each node needs to know which other nodes it should compress based on.

When there are more than two sensors in the cluster, each of them may be able to compress based on more than one node. However, allowing sensors to do so makes the control problem very complex. At the same time, as the energy spent when receiving is significant, if a node already compresses based on another node, it is likely to get very little gain when trying to receive and compress based on one more node. Therefore, we restrict our control schemes to those that satisfy the following constraint:

Constraint 6.5.1. *During each data-gathering round, each sensor is allowed to compress based on the data of at most one other sensor and that sensor must transmit uncompressed data.*

With the above constraint, we give the following definition for a CBC policy that controls the sensors during each data-gathering round.

Definition 6.5.2. *Let* $\mathbf{v} \subseteq \mathbf{N}$ *be a subset of the set of all* K *sensors, a* ***CBC policy*** *is a function* $\mu : \mathbf{v} \to \mathbf{v} \cup \{0\}$ *such that for* $i, k \in \mathbf{v}$, $\mu(k) = 0$ *if* k *is not allowed to compress based on any other node while* $\mu(k) = i$ *if* k *is allowed to compress based on* i. *Note that* $\mu(k) = i$ *implies* $\mu(i) = 0$.

Note that a particular CBC policy μ only controls the operation of those sensors belonging to $\mathbf{v}$, a subset of $\mathbf{N}$. This makes Definition 6.5.2 applicable even if not all K sensors in the cluster are active. It can be shown that, given a CBC policy μ, a transmission order can always be determined so that each node $k \in \mathbf{v}$ can carry out compression and transmission as specified by μ.

6.5.3 Control over Multiple Data-gathering Rounds

By definition, a particular CBC policy μ specifies how the sensors in the set $\mathbf{v} \subseteq \mathbf{N}$ operate during a particular data-gathering round. To control the sensors over multiple data-gathering rounds, we define a CBC scheme as:

Definition 6.5.3. *Let* $\mathbf{v} \subseteq \mathbf{N}$ *be a subset of the set of all* K *sensors, a* ***CBC scheme*** *is a policy-time set*

$$\mathbf{\Psi} = \Big\{(\mu_1, t_1), \dots (\mu_m, t_m)\Big\}$$

in which the pair (μ_i, t_i), $1 \leq i \leq m$, *indicates that CBC policy* μ_i *is employed on* $\mathbf{v}$ *for* t_i *data-gathering rounds. Furthermore, let* e_k^{res} *be the residual energy that node* k *has prior to the application of* $\mathbf{\Psi}$, *then* $\mathbf{\Psi}$ *is said to be* ***feasible*** *if and only if:*

$$\sum_{i=1}^{m} E_k(\mu_i(k)) t_i \leq e_k^{res}, \ \forall k \in \mathbf{v}. \tag{6.20}$$

Condition (6.20) guarantees that when $\mathbf{\Psi}$ is applied, no sensor in $\mathbf{v}$ consumes more than its residual energy.

6.5.4 Sensor Lifetime and System Performance

Let us suppose that some feasible CBC schemes are employed to control K sensors until all of them use up their energy and die. The operation of the cluster can be divided into K consecutive phases, with phase k starting when $k-1$ out of K sensors die and ending when k out of K sensors die. We then define a lifetime vector of the cluster as follows.

Definition 6.5.4. *The K- element vector L, with $L(k)$ being the time when phase k ends, is called a* ***lifetime vector*** *of the cluster. Furthermore, a lifetime vector L is said to be* ***achievable*** *if it is the result of the application of some K feasible CBC schemes, each controls one phase of the cluster operation.*

It is straightforward to prove the following lemma, which states that by applying the CBC approach, every node in the cluster will achieve at least the lifetime corresponding to the case when no node carry out joint data compression,

Lemma 6.5.5. *Let $\widetilde{L}$ be the lifetime vector achieved when no node carry out joint data compression, then for every achievable lifetime vector L, $L(k) \geq \widetilde{L}(k),\ \forall k \in \mathbf{N}$.*

Now, let us examine some options for characterizing the cluster data-gathering performance based on the lifetime vector L. For the most stringent performance, the cluster ceases functioning when one of its K sensors dies, i.e., at time $L(1)$.

For the least stringent case, we may assume that the cluster keeps on functioning until all of its sensors die, i.e., at time $L(K)$. However, in reality, when sensor nodes die one by one, what will be observed is a gradual decrease in the quality of the data-gathering job. The decrease here is in terms of information-fidelity and/or geographical coverage. This gradual decrease in performance can not be captured by any single element of the lifetime vector L. Therefore, we propose to maximize elements of L in sequence, with the maximization of the k^{th} element being carried out conditioned on the maximization of the 1^{st}, $2^{nd}, \ldots (k-1)^{th}$ elements. In a more concrete form, we adopt the following definition for the optimality of the cluster lifetime vector:

Definition 6.5.6. *An achievable lifetime vector L^* is said to be* ***optimal*** *if for every other achievable lifetime vector L, $L \neq L^*$, there exists $k \in \mathbf{N}$ such that*

$$L^*(i) \geq L(i), \quad \forall i \in \{1, \ldots k\}, \tag{6.21}$$

with at least one strict inequality.

Note that our optimality criteria gives priority to improving the lifetimes of nodes who die early. This will keep as many nodes to stay alive as possible, and therefore, assure a high-level data-gathering performance for a long period of time. This also leads to reduction in the variance among nodes' lifetimes, i.e., nodes die closer together.

6.6 Lifetime Vector Optimization Problem

Based on Definition 6.5.6, we introduce the following lifetime vector optimization (LVO) problem:

Lifetime Vector Optimization (LVO) Problem: *Given a cluster of K sensors, find K feasible CBC schemes that respectively control K phases of the cluster operation so that the resultant lifetime vector is optimal.*

6.6.1 A General Approach to Solve the LVO Problem

The LVO problem can be solved by the following K-step procedure.

- *Step* 1: Given all K sensors with their initial energies, we find a feasible CBC scheme that controls phase 1 of the cluster operation so that the time when one of the K sensors dies is maximized. Step 1 gives us L_1^* which is the maximum lifetime of the node who dies first.

- *Step* k, $2 \leq k \leq K$: The first $k-1$ steps give us $L_1^*, \ldots L_{k-1}^*$. Now the task is to find k feasible CBC schemes that control the first k phases of the cluster operation so that the time when i out of K sensors die is L_i^*, $\forall i < k$, and the time when k out of K sensors die is maximized. This conditional maximum time when k out of K sensors die is denoted by L_k^*.

Theorem 6.6.1. *The K feasible CBC schemes obtained in Step K solve the LVO problem.*

Proof. After Step K, we obtain K feasible CBC schemes that achieve the lifetime vector $(L_1^*,\ L_2^*, \ldots L_K^*)$. We will prove that this lifetime vector is optimal with respect to Definition 6.5.6.

Let L be any achievable lifetime vector and $L \neq (L_1^*,\ L_2^*, \ldots L_K^*)$. There must be k, $1 \leq k \leq K$, such that $L(i) = L_i^*$, $\forall i < k$ and $L(k) \neq L_k^*$. Note that L_k^* is the maximum time when k out of K sensors die, subject to the constraint that the time when i out of K sensors die is $L(i),\ \forall i < k$. Therefore, we must have $L(k) < L_k^*$. In other words, k satisfies the optimality condition in Definition 6.5.6 and $(L_1^*,\ L_2^*, \ldots L_K^*)$ is the optimal lifetime vector. $\square$

6.6.2 Linear Programming Formulation

Now let us show how each step in the K-step procedure described in Section 6.6.1 can be formulated as a linear programming (LP) problem. We do so for Step 1 and 2. For Steps $k,\ k > 2$, the formulation is similar.

Formulating Step 1 as an LP

As the number of CBC policies in phase 1 can be very large, what we will do first is to narrow down the policies that should be time-shared. Given a CBC policy μ, let us denote by $\mathbf{u}$ the set of all nodes that transmit without compressing based on another node. We must have:

$$\forall k \in \mathbf{N} \backslash \mathbf{u},\ \mathbf{u} \cap \mathbf{N}_k \neq \emptyset. \tag{6.22}$$

In other words, each node in $\mathbf{N}\backslash\mathbf{u}$ must be able to receive the transmission of at least one node in $\mathbf{u}$. Furthermore, we only need to consider those policies μ that satisfy:

$$\mu(k) = \gamma_1^k(\mathbf{u}) = \begin{cases} 0, \quad \forall k \in \mathbf{u}, \\ \underset{j \in \mathbf{u} \cap \mathbf{N}_k}{\arg\min} \ \{E_k(j)\},\ \forall k \in \mathbf{N}\backslash\mathbf{u}. \end{cases} \tag{6.23}$$

This is because given a set $\mathbf{u}$ of nodes that transmit without compressing, all other nodes should choose to compress based on the node that result in the most energy saving. As a result, each policy μ that we will time-share is completely specified if the set of nodes that transmit without compressing is given.

Let $\mathcal{U}_1$ be the set of all subsets of $\mathbf{N}$ that satisfies condition (6.22), i.e.,

$$\mathcal{U}_1 = \{\mathbf{u} \subseteq \mathbf{N} \mid \forall k \in \mathbf{N} \backslash \mathbf{u},\ \mathbf{u} \cap \mathbf{N}_k \neq \emptyset\}. \tag{6.24}$$

Also, let $t_1^{\mathbf{u}}, \mathbf{u} \in \mathcal{U}_1$, be the number of data-gathering rounds that all nodes belonging to $\mathbf{u}$ transmit without compressing while all nodes not belonging to $\mathbf{u}$ carry out data compression. Note that the subscript '$_1$' of γ_1^k, $\mathcal{U}_1$, and $t_1^{\mathbf{u}}$ is used to indicate that these are function or parameters of phase 1. As we have mentioned, when the lifetimes of sensors are much longer than each data-gathering rounds, $t_1^{\mathbf{u}}$ can be treated as real variables. Then, the problem of maximizing the lifetime of the node who dies first can be written as the following linear program:

$$\underset{t_1^{\mathbf{u}},\ \forall \mathbf{u} \in \mathcal{U}_1}{\arg\max} \sum_{\forall \mathbf{u} \in \mathcal{U}_1} t_1^{\mathbf{u}} \tag{6.25}$$

subject to:

$$t_1^{\mathbf{u}} \geq 0, \quad \forall \mathbf{u} \in \mathcal{U}_1, \tag{6.26}$$

$$\sum_{\forall \mathbf{u} \in \mathcal{U}_1} \left(t_1^{\mathbf{u}} E_k(\gamma_1^k(\mathbf{u})) \right) \ \leq \ e_k, \quad \forall k \in \mathbf{N}. \tag{6.27}$$

Solving the above LP gives us a CBC scheme that maximizes the time until at least one of the K sensors die. This maximum lifetime is denoted by L_1^*. For this particular CBC scheme, let us denote by $\mathcal{D}_1$ the set of nodes that actually die at time L_1^*. $\mathcal{D}_1$ can be determined just by checking the residual energies of all K sensors after phase 1.

Formulating Step 2 as an LP

Let us first consider the case when the set $\mathcal{D}_1$, obtained by solving the LP for Step 1, only has one element, denoted by k^*. In other words, only sensor k^* dies at time L_1^*. In phase 2 we are left with $K-1$ nodes in the set $\mathbf{N}\backslash\{k^*\}$. Now, similar to γ_1^k, $\mathcal{U}_1$, $t_1^{\mathbf{u}}$ ($\forall \mathbf{u} \in \mathcal{U}_1$) of phase 1, we can define γ_2^k, $\mathcal{U}_1$, $t_2^{\mathbf{v}}$ ($\forall \mathbf{v} \in \mathcal{U}_2$) for phase 2. Then the task of Step 2, i.e., to find two CBC schemes that control phases 1 and 2 so that the duration of phase 1 is L_1^* and the duration of phase 2 is maximized, can be written as the following LP.

$$\underset{t_1^{\mathbf{u}},\ t_2^{\mathbf{v}},\ \forall \mathbf{u}\in\mathcal{U}_1,\ \forall \mathbf{v}\in\mathcal{U}_2}{\arg\max} \sum_{\forall \mathbf{v}\in\mathcal{U}_2} t_2^{\mathbf{v}} \tag{6.28}$$

subject to:

$$t_1^{\mathbf{u}} \geq 0 \ \forall \mathbf{u} \in \mathcal{U}_1, \quad t_2^{\mathbf{v}} \geq 0 \ \forall \mathbf{v} \in \mathcal{U}_2, \tag{6.29}$$

$$\sum_{\forall \mathbf{u}\in\mathcal{U}_1} t_1^{\mathbf{u}} \ = \ L_1^*, \tag{6.30}$$

$$\sum_{\forall \mathbf{u}\in\mathcal{U}_1} \Big(t_1^{\mathbf{u}} E_{k^*}(\gamma_1^{k^*}(\mathbf{u}))\Big) \ \leq \ e_{k^*}, \tag{6.31}$$

$$\sum_{\forall \mathbf{u}\in\mathcal{U}_1} \Big(t_1^{\mathbf{u}} E_k(\gamma_1^k(\mathbf{u}))\Big) + \sum_{\forall \mathbf{v}\in\mathcal{U}_2} \Big(t_2^{\mathbf{v}} E_k(\gamma_2^k(\mathbf{v}))\Big) \ \leq \ e_k, \quad \forall k \in \mathbf{N}\backslash\{k^*\}. \tag{6.32}$$

In the cases when $\mathcal{D}_1$ contains more than one node, we will need to formulate the above LP for each possible value of k^*, and determine which one leads to the maximum lifetime of phase 2.

3) Complexity of Solving Each Step by LP

Note that of all K steps, Step 1 involves solving the smallest LP. The size of the LP for Step 1 depends on the cardinality of the set $\mathcal{U}_1$, which in turn depends on the cluster topology and sensor's energy model. In the worst case,

$\mathcal{U}_1$ contains all non-empty subsets of $\mathbf{N}$, and therefore, has the cardinality of $2^K - 1$. This means it is only practical to solve the above LPs when the number of nodes in the cluster is small.

6.7 Heuristic Algorithm

In this section, we propose heuristic CBC schemes which can be obtained at a much lower complexity compared to solving the linear programming problems in Section 6.6.2. In Section 6.9, we will present numerical results which show that the heuristic schemes achieve a near optimal lifetime vector. We will focus on the heuristic scheme that controls phase 1 of the cluster operation. The schemes for other phases can be constructed in a similar manner.

In the heuristic CBC scheme that controls phase 1, i.e., $\{(\mu_1, t_1), \ldots (\mu_m, t_m)\}$, each policy μ_i is employed for an interval of T data-gathering rounds, i.e., $t_i = T,\ i = 1, \ldots m$, where T is a fixed integer. For each interval, a CBC policy is selected in a greedy way, with the objective of maximizing the minimum residual energy of K nodes after the interval.

6.7.1 A CBC Policy for T Data-gathering Rounds

Let interval $n,\ n = 1, 2, \ldots,$ denote the time from the beginning of data-gathering round $(n-1)T + 1$ until the end of data-gathering round nT. Let $e_k^n,\ e_k^n \geq 0,\ k \in \mathbf{N}$, be the residual energy of node k at the beginning of interval n. Also, let μ_n be the CBC policy being employed in interval n. If no node uses up its energy during interval n, the residual energy of node k at the beginning of interval $n+1$ is

$$e_k^{n+1} = e_k^n - TE_k(\mu_n(k)). \tag{6.33}$$

The lifetime of each node is directly related to its residual energy. Therefore, during each interval, it is intuitive to employ a greedy CBC policy that maximizes the minimum value of the residual energy of all K nodes after the interval. In other words, for interval n, we will find the CBC policy μ_n^* that satisfies

$$\mu_n^* = \arg\max_{\mu_n} \left\{ \min_{k \in \mathbf{N}} \{e_k^{n+1}\} \right\}. \tag{6.34}$$

In order to obtain μ_n^*, we start with a CBC policy μ in which no node compresses based on any other node and improve μ in each iteration. Policy μ is improved by first identifying the node i^* that will have the least residual energy at the end of interval n if policy μ is applied and then let i^* compress based on another node. When there are more than one node that i^* can compress based on, i^* will choose the node j^* that satisfies

$$j^* = \arg\max_{j \in \mathbf{Q}} \left\{ \min\{e_{i^*}^n - TE_{i^*}(j), \left(e_j^n - TE_j(0)\right) / \mathbb{1}(j \notin \mathbf{U})\} \right\}. \tag{6.35}$$

The reason for i^* to compress based on j^* selected by (6.35) is that if we let i^* compress based on some node j, then j is not allowed to compress based on any node, and j can become the node who has the least residual energy at the end of interval n. In (6.35), $\mathbf{Q}$ is the set of nodes that i^* can compress based on while $\mathbf{U}$ consists of nodes that are not able to compress and nodes that have already been used by other nodes for their data compression. After improving $e_{i^*}^{n+1}$ by letting i^* compress based on j^*, we move to the next iteration and repeat the process.

We name the above algorithm **Single_CBC** and give its pseudo-code in Fig. 6.6. Note that the inputs for algorithm Single_CBC are the residual energies of the K sensors at the beginning of interval n, i.e., $(e_1^n, \dots e_K^n)$. The output of

Single_CBC is a CBC policy that controls K sensors during interval n. As has been mentioned, **U** denotes the set of nodes that are either used by other nodes for their data compression and/or not able to compress. Besides, **V** denotes the set of nodes who compress based on some nodes in **U**.

6.7.2 A Heuristic CBC Scheme for Phase 1

By repeatedly applying the Single_CBC algorithm until one of the sensor nodes uses up its energy and dies, we obtain a set of CBC policies, each control the collaboration of K sensors for T data-gathering rounds. We name the algorithm that does so the **Multiple_CBC**. The inputs for Multiple_CBC are the initial energy of K nodes, i.e., $(e_1, \ldots e_K)$. Multiple_CBC outputs a sequence of CBC policies that are employed until one of the sensors dies. The pseudo-code for Multiple_CBC is presented in Fig. 6.7.

6.7.3 Complexity of Heuristic Algorithm

Heuristic CBC schemes for controlling phases $2, 3, \ldots K$ can be obtained in a similar manner to that of phase 1. As the complexity for obtaining heuristic CBC schemes is highest for phase 1, let us determine this complexity.

From the pseudo-code of Single_CBC, it can be seen that the main tasks inside the loop are to find i^* and j^*. Both of these involve finding the minimum value from a set of at most K elements and therefore, the complexity is of the order $O(K)$. At the same time, the main loop is repeated for no more than K times. Therefore, we can conclude that the worst-case complexity of Single_CBC is $O(K^2)$.

During each iteration of Multiple_CBC algorithm, Single_CBC algorithm is

Algorithm: **Single_CBC**$(e_1^n, \dots e_K^n)$

$\mu(k) \leftarrow 0, \ \forall k \in \mathbf{N}$

$\mathbf{U} \leftarrow \emptyset; \quad \mathbf{V} \leftarrow \emptyset$

loop

$$i^* \leftarrow \underset{k \,\in\, \mathbf{N}\backslash(\mathbf{U}\cup\mathbf{V})}{\arg\min} \{e_k^n - TE_k(\mu(k))\}$$

$$\mathbf{Q} \leftarrow \Big\{k \in \mathbf{N}_{i^*}\backslash\mathbf{V} \mid E_{i^*}(k) < E_{i^*}(0)\Big\}$$

if $\mathbf{Q} \neq \emptyset$

$$j^* \leftarrow \underset{j\in\mathbf{Q}}{\arg\max} \Big\{ \min\{e_{i^*}^n - TE_{i^*}(j), (e_j^n - TE_j(0))/\mathbb{1}(j \notin \mathbf{U})\}\Big\}$$

$\mu(i^*) \leftarrow j^*$

$\mathbf{U} \leftarrow \mathbf{U} \cup \{j^*\}; \quad \mathbf{V} \leftarrow \mathbf{V} \cup \{i^*\}$

else

$\mathbf{U} \leftarrow \mathbf{U} \cup \{i^*\}$

endif

if $\mathbf{U} \cup \mathbf{V} = \mathbf{N}$

break

endif

endloop

return μ

Figure 6.6: Pseudo-code of algorithm **Single_CBC**$(e_1^n, \dots e_K^n)$. Inputs are the residual energies of K sensors at the beginning of interval n, i.e., $(e_1^n, \dots e_K^n)$. Output is CBC policy μ that will be used to control K sensors during interval n.

Algorithm: Multiple_CBC$(e_1, \dots e_K)$

$\mathbf{\Psi} \leftarrow [\,]$

$e_k^{res} \leftarrow e_k, \ \forall k \in \mathbf{N}$

loop

$\mu \leftarrow$ **Single_CBC**$(e_1^{res}, \dots e_K^{res})$

$\mathbf{\Psi} \leftarrow [\mathbf{\Psi}, \ \mu]$

$e_k^{res} \leftarrow e_k^{res} - TE_k(\mu(k)), \ \forall k \in \mathbf{N}$

if $\min_{k \in \mathbf{N}}\{e_k^{res}\} \ \leq \ 0$

break

endif

endloop

return $\mathbf{\Psi}$

Figure 6.7: Pseudo-code of algorithm **Multiple_CBC**$(e_1, \dots e_K)$. Inputs are the initial energies of K sensors, i.e., $(e_1, \dots e_K)$. Output is a sequence of CBC policies, each policy is employed to control one interval of T rounds.

carried out. The number of iterations being taken in Multiple_CBC depends on the lifetime of the node who dies first. We note that the energy consumed by each node in a data-gathering round is lower-bounded by the energy consumed in the electronic circuits. In particular, if in each data-gathering round, each node is required to communicate a packet of length R bits (without compression) to the cluster head, then no matter whether a node compresses based on other nodes or not, the energy consumed in each round is lower bounded by $E_{lb} = E_e R$. Therefore, the lifetime of sensor k, $k = 1, \ldots K$, is upper-bounded by $L_{ub} = \frac{e_k}{E_{lb}}$. As the upper-bound L_{ub} does not grow with K, the number of iterations of Multiple_CBC algorithm does not grow with K either. As a result, the complexity of Multiple_CBC algorithm is of the same order of that of the Single_CBC algorithm, which is equal $O(K^2)$.

6.8 Reflections on the CBC Approach

6.8.1 Startup Cost of Sensor Nodes

For wireless sensors, startup cost refers to the energy consumed during the radio startup transient [SCI+01], [RSPS02]. Note that no data can be transmitted or received during this transient phase. One way to minimize the negative effect of this (wasted) energy is to operate at a large packet size so that the total energy consumed by the transceiver unit is dominated by transmission and reception energy [SCI+01].

In our CBC schemes, when a node wants to receive and compress based on the data of another node, its radio needs to be active during at least two time slots in each data-gathering round, i.e., one is for receiving and the other is for transmitting data. If these receiving and transmitting time slots are not adjacent to each other, in order to conserve energy, the node may need to turn off the radio component after the receiving and then turn it on again for transmission. Doing so will not cause any problem as long as the radio startup cost is negligible.

For the case when the startup cost is significant, we can mitigate the problem of non-adjacent receiving and transmitting time slots by constraining that in each CBC policy, at most one node can compress based on any particular node. This will allow a node to transmit right after receiving and compressing its data. Note that the constraint can be easily incorporated into our linear programming and heuristic approaches in Sections 6.6.2 and 6.7. In Section 6.9, we will present numerical result to show that with this extra constraint, our CBC schemes still yield a significant improvement for sensors' lifetimes.

6.8.2 Packet Transmission Errors

So far, when studying the CBC approach, we have assumed that the packet loss due to transmission errors is negligible. Now let us consider how our CBC schemes perform when packet transmission errors are taken into account.

We suppose that, in a particular CBC scheme, sensor k is assigned to compress based on sensor i during some time interval. This will improve the lifetime of k. However, due to transmission errors, in some data-gathering rounds, k may not be able to receive packets sent by i and therefore, can not compress its data. As a result, our CBC schemes will achieve less lifetime improvement, relative to the case when all transmissions are successful.

Still referring to the above scenario, we assume that k actually receives a packet sent by i and used that to compress its own packet. However, let us suppose that the packet of i is not received successfully by the cluster head H. If H keeps on requesting i to resend its packet until a successful reception, then the compressed packet of k will eventually be decoded. On the other hand, if no retransmission is allowed, the loss of the packet of i will leads to the loss of the packet of k as this packet can not be decompressed. As a result, under our CBC schemes, those nodes who compress based on others' data can incur a higher packet loss probability.

For node k, the packet loss probability will be worst when the packet loss processes corresponding to the transmission from i to H and the transmission from k to H are independent. In that case, let P_i^e and P_k^e be the packet loss probabilities for the transmissions from i and k (to H) respectively, the packet loss probability for k can be written as:

$$P_{k|i}^e = P_k^e + P_i^e - P_k^e P_i^e \approx P_k^e + P_i^e. \tag{6.36}$$

As our CBC schemes may increase the packet loss rate for nodes that compresses based on others, apart from the lifetime improvement, it is useful to look at the performance in terms of the total number of packets successfully transmitted by each node throughout its lifetime. In Section 6.9 we will present result to show that even with a high packet loss rate (10%), our CBC schemes still result in significant increases in the total number of successful packets transmitted by each node.

6.8.3 Effects on the Relaying Network

Now, let us discuss the effects that our CBC approach can have on the relaying network formed by type II nodes. First of all, as nodes in each cluster jointly compress their data, the amount of data sent to the cluster heads will be reduced. This can allow the cluster heads to spend less energy receiving. Secondly, as nodes encode their data based on explicit side information, the decoding scheme at each cluster head will not be complex. In fact, the cluster heads may not want to decompress the data, since they will eventually perform data fusion/aggregation. Finally, after data fusion/aggregation, there will be no change on the amount of data flowing out of each cluster. Therefore, other parts of the relaying network are not affected by the data compression carried out within each cluster.

Based on the above discussion, we state that our CBC approach is independent to the operation of the relaying network. Therefore, it can be applied in conjunction with energy-efficient routing schemes that have been proposed for WSNs ([IGE00]).

6.9 Numerical Study

In this section, we present numerical results which show the performance gain, i.e., the increase in sensors' lifetimes and number of packets successfully transmitted, when the CBC approach is employed. We will compare the performance of three control schemes, i.e., the optimal scheme obtained by solving the LPs formulated in Section 6.6.2, the heuristic scheme proposed in Section 6.7, and finally the scheme in which all sensors transmit to the cluster head without joint compression.

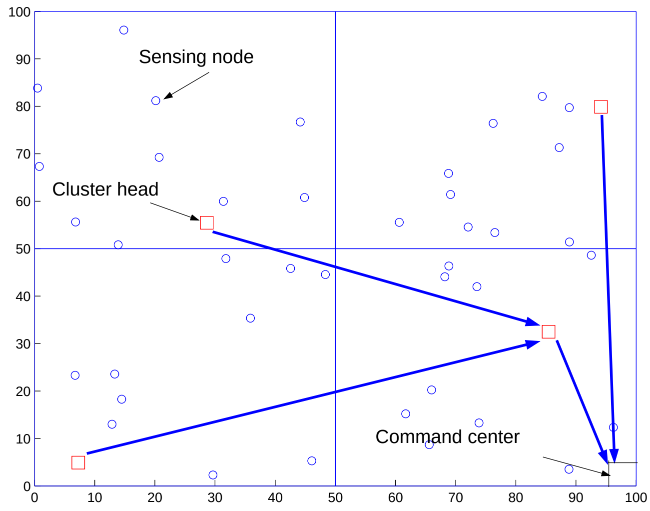

Figure 6.8: An example of a network of size $100 \times 100m$. The monitoring area is divided into four clusters. In each cluster, there are $K = 10$ sensing nodes and one cluster head. Sensing nodes and cluster heads are deployed randomly and uniformly within their cluster area.

6.9.1 Experimental Model

The monitored field is represented by a square area of size D meters. This area is further divided into C^2 disjoint clusters, each is a square of size $\frac{D}{C}$. In each cluster, there are K sensors and one cluster head. We assume that the sensor nodes, together with the cluster head, are deployed randomly within each cluster, with their coordinates uniformly distributed. In Fig. 6.8, a sample network of size $D = 100$ meters, divided into four clusters and with $K = 10$ sensors per cluster, is shown.

The energy model of each sensor node is as described in Section 6.3.3 with: $E_a = 100\text{pJ/bit/m}^2$, $E_c = 5\text{nJ/bit}$, and $E_e = 10,\ 50,\ 100\text{nJ/bit}$. Each sensor node has an initial energy storage of 5J. In each round, without compression, each sensor needs to send a packet of length $R = 400$ bits to the cluster head. For the sake of simplicity, we ignore the bits used in the packet header. This assumption is justified when the size of the packet header is much smaller than that of the data payload. We also assume that, if a particular node k compresses based on another node, then the compression ratio, i.e., $\frac{r}{R}$, is fixed.

Each of the results presented in the following section is obtained by generating 500 instances of the network and averaging the performance of tested schemes. Note that $L(k),\ k = 1, \ldots K$, denotes the time when k out of the K sensors in a cluster die when some control scheme is employed.

6.9.2 Results and Discussion

In Fig. 6.9, we show the percentage increases in sensors' lifetimes when the optimal control schemes and the heuristic control schemes are applied, relative to when no node compresses its data. The percentage increases in the lifetimes of nodes who die first, second, fifth, and tenth are plotted versus the compression ratio $\frac{r}{R}$. Here, we assume that all packets are successfully transmitted. As can be seen, the lifetime improvements strongly depend on the compression ratio $\frac{r}{R}$, i.e., on the spatial correlation among data collected at different sensors. When $\frac{r}{R}$ is low, the performance gain of both optimal and heuristic schemes are very significant. The gain is largest for $L(1)$ while there is negligible gain for $L(10)$. This is exactly what our objective is; we want to improve the lifetimes of those nodes who die earlier than others. Another important observation is that the performance of the heuristic scheme is nearly the same as that of the optimal control scheme. This indicates that we can use the heuristic scheme, which has much lower complexity without sacrificing performance.

We then look at how the performance of the heuristic scheme depends on the number of sensors per cluster and the cluster size. In Fig. 6.10 we show the

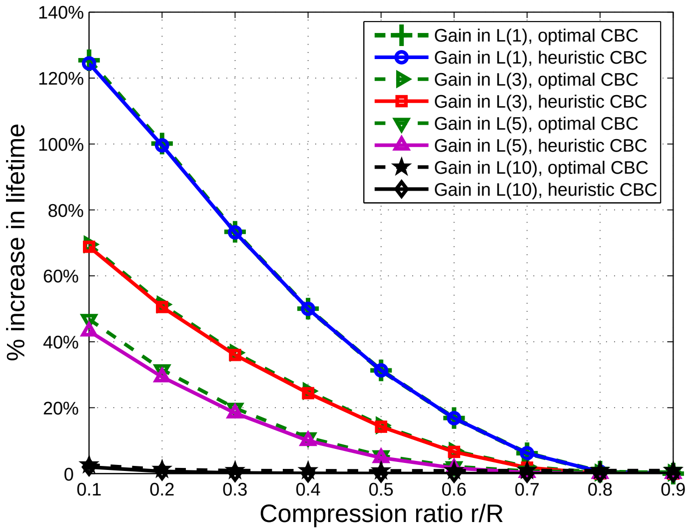

Figure 6.9: Percentage increases (relative to no compression) in sensors' lifetimes versus compression ratio when the optimal CBC and heuristic CBC schemes are applied. $L(1)$, $L(3)$, $L(5)$, $L(10)$ are the lifetimes of nodes who die first, third, fifth, and tenth, respectively. There are $K = 10$ nodes in each cluster and the energy model is: $E_a = 100\text{pJ/bit/m}^2, E_e = 50\text{nJ/bit}$ and $E_c = 5\text{nJ/bit}$. Packet loss is assumed to be negligible.

percentage increase for $L(1)$ when the heuristic scheme is applied, as compared to the case when no node carries out compression for different values of K and D, i.e., $K = 10,\ 25$ and $D = 100\text{m},\ 200\text{m}$. The percentage increase is plotted against the compression ratio. As can be seen, the gain in lifetime increases in the number of nodes per cluster. This can be explained by the fact that, when there are more nodes in each cluster, the distance among them gets shorter, each node has more options on which node it can used to compress its data. At the same time, when the cluster size D is increased, the performance gain also increases. This is because with a larger cluster size, the average distance from sensors to the cluster head increases and in Section 6.4.2, we have shown that this increase in the distance will give node more incentive to jointly compress

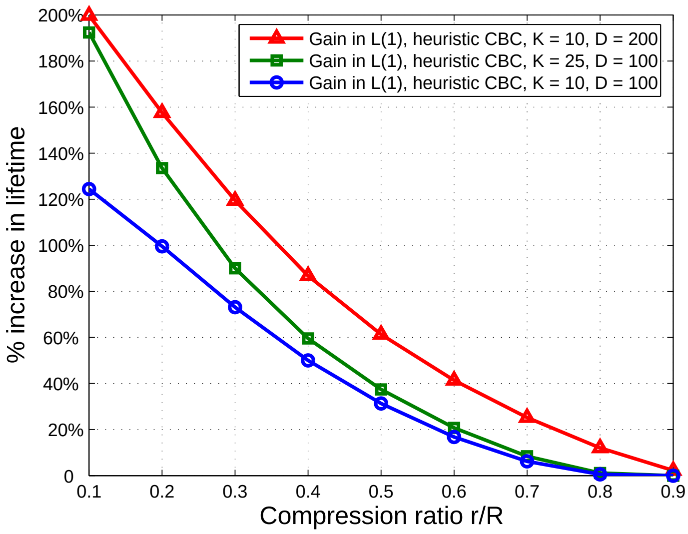

Figure 6.10: Percentage increase (relative to no compression) in the lifetime of the node who dies first versus compression ratio when the heuristic CBC scheme is applied. The cluster size is $D = \{100, 200\ m\}$ and the number of sensors/cluster is $K = \{10, 25\}$. The energy model is: $E_a = 100$pJ/bit/m^2, $E_e = 50$nJ/bit and $E_c = 5$nJ/bit. Packet loss is assumed to be negligible.

data.

Next, we look at how the performance gain of the heuristic scheme depends on the energy model of sensor nodes. In particular, we let the value of electronic energy, i.e., E_e vary from 10 to 100nJ/bit while still keeping the amplifier and processing energy unchanged. In Fig. 6.11, we plot the percentage increase for the lifetime of the node who dies first versus the compression ratio for $E_e =$ 10, 50 and 100nJ/bit. As expected, the gain decreases in E_e. However, even when $E_e = 100$nJ/bit, the gain is still about 30% for the compression ratio of 0.5.

In Section 6.8.1, we suggest that to deal with the scenario when the radio startup cost is significant, an extra constraint, i.e., no more than one node can

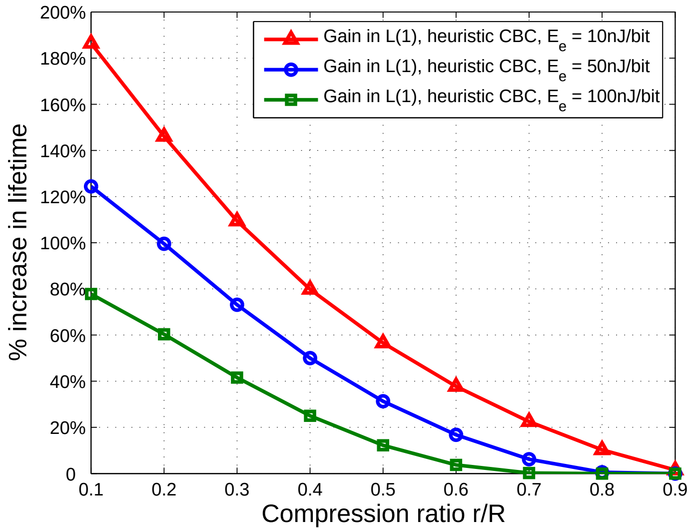

Figure 6.11: Percentage increase (relative to no compression) in the lifetime of the node who dies first versus compression ratio when the heuristic CBC scheme is applied. There are $K = 10$ nodes in each cluster and the energy model is: $E_a = 100\text{pJ/bit/m}^2$, $E_c = 5\text{nJ/bit}$ and E_e takes the values $\{10, 50, 100\text{nJ/bit}\}$. Packet loss is assumed to be negligible.

compress based on any node can be enforced. In Fig. 6.12, we look at the performance of the heuristic scheme with this extra constraint. Here we plot the percentage increase in different components of the lifetime vector L. It is obvious that by enforcing the extra constraint, the increase in sensor lifetime is less, however, as can be seen in Fig. 6.12, the gain in $L(1)$ is still very significant. In particular, when the compression ratio is 0.5, applying the modified heuristic scheme results in 30% increase in the lifetime of the node who dies first.

Finally, let us look at how our heuristic CBC schemes perform under packet loss due to transmission errors. Here, we assume that the packet loss processes for the transmissions between different pairs of nodes in the cluster are independent (note that this is the worst case assumption) and with the same packet

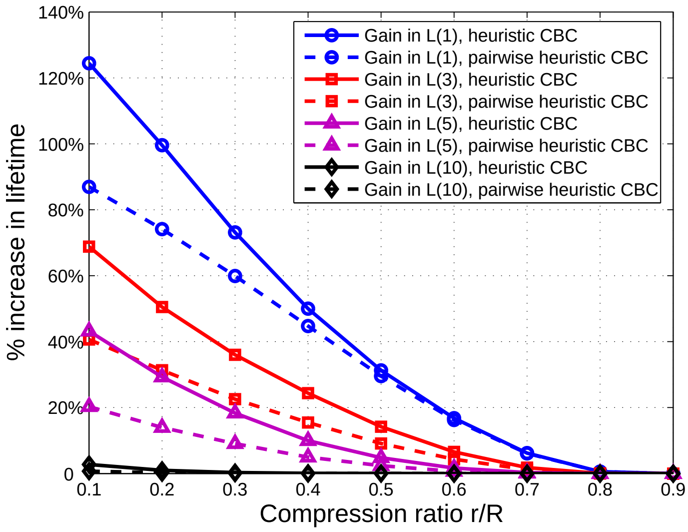

Figure 6.12: Percentage increases in $L(1)$, $L(3)$, $L(5)$, $L(10)$ versus compression ratio when the heuristic CBC and pairwise heuristic CBC schemes are applied. Pairwise heuristic CBC schemes alow at most one node to compress based on any particular node. There are $K = 10$ nodes in each cluster and the energy model is: $E_a = 100\text{pJ/bit/m}^2$, $E_e = 50\text{nJ/bit}$ and $E_c = 5\text{nJ/bit}$. Packet loss is assumed to be negligible.

loss probability denoted by P^e. In Fig. 6.13, we plot the performance of our heuristic CBC schemes, in terms of the percentage increase in the total number of packet successfully transmitted for the node who dies first. Different packet loss probabilities are used, i.e., $P^e = 0,\ 1\%,\ 5\%,\ 10\%$. As can be seen, even when the packet loss probability is relatively high, i.e., at 10%, the performance gain for the node who dies first is still very significant. This suggests that our CBC approach is robust against packet loss due to transmission errors. Note also in Fig. 6.13 that the points at which performance curves cut the zero-level line is where the CBC approach does not give any performance improvement.

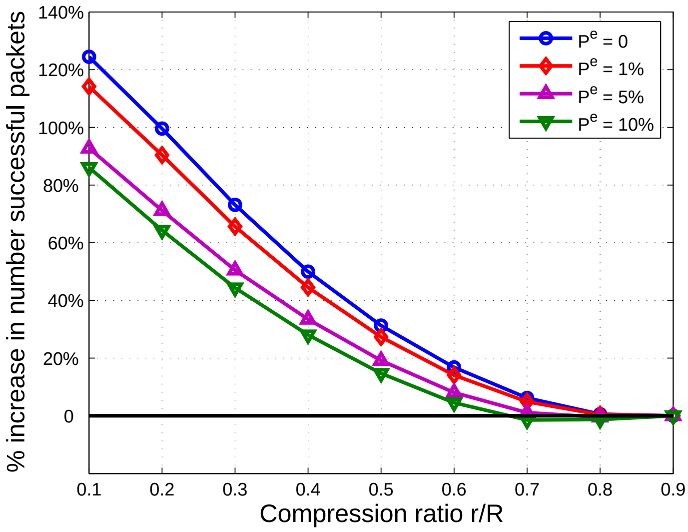

Figure 6.13: Percentage increases in the number of packets successfully transmitted for the node who dies first when the heuristic CBC scheme is applied. There are $K = 10$ nodes in each cluster and the energy model is: $E_a = 100\text{pJ/bit/m}^2, E_e = 50\text{nJ/bit}$ and $E_c = 5\text{nJ/bit}$. Packet loss processes of different transmissions are independent and with the same packet loss probability P^e. The points at which performance curves cut the zero-level line is where the CBC approach does not give any performance improvement.

6.10 Conclusion

In this chapter, we proposed a novel approach in which the inherent broadcast nature of the wireless medium is exploited by sensor nodes to carry out joint data compression and conserve energy. This is different from the usual abstraction of a communication network by a communication graph, in which nodes interact in a point-to-point fashion.

Our metric of interest is sensor network lifetime. We first presented algorithms which optimize the lifetime vector of the network, meaning that any other algorithm will not increase the lifetime of the node which dies first. We then proposed a heuristic algorithm which has significantly lower computational complexity with near optimal performance. Important characteristics of wireless sensor networks such as node startup cost and packet loss due to transmission errors are also considered. Extensive numerical results are presented to support our approach.

Taking a broader view, our work in this chapter highlights two important issues in designing WSNs. Firstly, it is important to not over simplify the network model when designing energy-constrained WSNs. One example of over-simplification, we believe, is the popular point-to-point link abstraction of the wireless medium. Secondly, our results show the importance of exploiting the opportunity to collaborate in wireless settings. This is because on one hand, different nodes in a WSN experience different performance/resource constraints while on the other hand, what is important is not a sensor's individual performance, but rather the network's collective performance.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this thesis, we have studied different cross-layer scheduling and transmission schemes for energy-constrained wireless systems. Our objective is to achieve good system performance while conserving nodes' energy. Let us wrap up by summarizing the main contributions of this thesis and discussing some avenues for future research.

We started by considering cross-layer adaptive transmission for a time-slotted, single-user system with stochastic data arrival, finite-length buffer, and time-varying wireless channel. The objective is to adapt the transmission power and rate according to the buffer and channel conditions in order to maximize the system throughput, subject to an average transmit power constraint. In Chapters 3 and 4, we have studied this buffer and channel adaptive transmission problems under various scenarios, i.e., with complete and incomplete system state information, and with or without a bit-error-rate constraint. In all the cases, we showed how good buffer and channel adaptive transmission schemes can be obtained and provided numerical results for their performance. An interesting structural property of optimal cross-layer adaptive transmission policies have also been identified.

Next, in Chapter 5, we studied a multiple-access scenario in which multiple users transmit data to a center node over a shared time-varying wireless channel. Two control decisions are made in each time slot, i.e., a scheduling decision which assigns the channel to one of the users, and a transmission decision which sets the transmit power and rate. All scheduling/transmission policies employed must satisfy the average transmit power constraint of each node. We solved the problem of finding optimal cross-layer adaptive scheduling/transmission policy which adapts to the buffer and channel conditions of all users so that the total system throughput is maximized. We then used the performance of this optimal policy as a benchmark to assess the performance of simpler cross-layer adaptive scheduling/transmission schemes. This allowed us to draw useful guidelines for controlling energy-constrained multiple-access systems.

Based on the results in Chapters 3, 4, and 5, we can draw the following general conclusions about cross-layer adaptation for physical and MAC layers. First, when nodes are equipped with limited batteries/buffers and operate within a dynamic environment, cross-layer design is essential to achieve good system performance. Second, when statistics of multiple layers are taken into account, popular intuitions associated with layer design may no longer hold. For example, as it has been shown in Chapter 3, under certain conditions, the structure of the throughput-optimal adaptive transmission policies can be the reverse of water-filling. Third, our results in Chapters 3, 4, and 5 demonstrated that cross-layer adaptation can be applied in a wide range of system scenarios.

A challenging extension to the adaptive scheduling/transmission problem in Chapter 5 is to consider cases when multiple users can simultaneously transmit on the same channel. Then, interference needs to be taken into account. For such cases, it is very complex to obtain optimal joint adaptive scheduling/transmission policies and it is more reasonable to look for sub-optimal policies that perform well and can be obtained at lower complexity. We reserve this problem for future research.

In Chapter 6, we studied a problem of combining scheduling, transmission, and data compression to conserve energy in spatially correlated sensor networks.

We considered a class of cluster-based wireless sensor networks with periodic data gathering. Since wireless transmission is inherently broadcast, when one sensor node transmits data to the cluster head, other nodes in its coverage area can receive the transmitted data. When data collected by different sensors are correlated, each sensor can utilize the data it overhears from other sensors to compress its own data and conserve energy in its own transmissions. Based on this observation, we formulated a problem in which sensors in each cluster are scheduled to transmit to the cluster head so that they can collaborate in joint source compression in order to maximize the network lifetime. We showed that this lifetime optimization problem can be solved by a sequence of linear programming problems. We also proposed a heuristic scheme, which has low complexity and achieves near optimal performance.

With respect to the studies of cross-layer adaptive scheduling/transmission in Chapters 3, 4, 5, the problem in Chapter 6 demonstrated that cross-layer design can still be beneficial even when there is no variation in the system parameters to adapt to.

The cross-layer work in Chapter 6 highlighted two important issues in designing wireless sensor networks. Firstly, it is important to not over simplify the network model when designing energy-constrained wireless sensor networks. One example of over-simplification, we believe, is the popular point-to-point link abstraction of the wireless medium. Secondly, our results show the importance of exploiting the opportunity to collaborate in wireless settings. This is because on one hand, different nodes in a WSN experience different performance/resource constraints while on the other hand, what is important is not a sensor's individual performance, but rather the network's collective performance.

There are several future directions for the work in Chapter 6. First, it will be useful to obtain distributed algorithms that exploit the wireless broadcast property to conserve sensors' energy. Second, we can study a similar problem for non-clustered sensor networks. Finally, the idea of exploiting wireless broadcast advantage can be applied at higher network layers, for example, in designing energy-efficient and reliable routing algorithms.

APPENDIX A

PROOF OF LEMMA 3.3.1

Before proving Lemma 3.3.1, let us prove the following Lemmas A.0.1, A.0.2, and A.0.3.

Lemma A.0.1. *For all* $0 \le g < K$, $J^*_\alpha(b, g)$ *is increasing in the buffer occupancy* b.

Proof. This lemma can be proved by induction. First of all, let J_0 be a bounded and increasing function on the state space (b, g). For $i = 1, 2, \ldots$, let

$$\begin{aligned} J_i(b, g) = \min_u \Big\{ & C_I(b, g, u) \\ & + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g, g') \times p_A(a) \times J_{i-1}\big(q(b-u, a), g'\big) \Big\}. \end{aligned} \tag{A.1}$$

Note that from the value iteration algorithm for solving discounted cost problem (3.23), we have $J^*_\alpha(b, g) = \lim_{i \to \infty} J_i(b, g)$ for all $0 \le b \le B$ and $0 \le g < K$. Now assuming $J_{i-1}(b, g)$ is increasing in b for all g, we will show that $J_i(b, g)$ is also increasing in b for all g. Then by induction, $J^*_\alpha(b, g)$ is increasing in b for all g.

For $0 < b \le B$, let u^* be the value that achieve the minimization in (A.1), we consider the following two possibilities.

a) $u^* = 0$

From (A.1) we have:

$$\begin{aligned} J_i(b-1, g) = \min_u \Big\{ & C_I(b-1, g, u) \\ & + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g, g') \times p_A(a) \times J_{i-1}\big(q(b-u-1, a), g'\big) \Big\} \\ \le\ & C_I(b-1, g, 0) + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g, g') \times p_A(a) \times J_{i-1}\big(q(b-1, a), g'\big). \end{aligned} \tag{A.2}$$

As $C_I(b-1,g,0) \le C(b,g,0)$ and $J_{i-1}\big(q(b-u-1,a),g'\big) \le J_{i-1}\big(q(b-u,a),g'\big)$ from (A.2) we have:

$$\begin{aligned} & J_i(b-1,g) \\ & < C_I(b,g,0) + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g,g') \times p_A(a) \times J_{i-1}\big(q(b,a),g'\big) \\ & = J_i(b,g). \end{aligned} \tag{A.3}$$

b) $u^* > 0$

From (A.1) we have:

$$\begin{aligned} & J_i(b-1,g) \\ & = \min_u \Big\{ C_I(b-1,g,u) \\ & \quad + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g,g') \times p_A(a) \times J_{i-1}\big(q(b-u-1,a),g'\big) \Big\} \\ & \le C_I(b-1,g,u^*-1) + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g,g') \times p_A(a) \times J_{i-1}\big(q(b-u^*,a),g'\big) \\ & < C_I(b,g,u^*) + \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g,g') \times p_A(a) \times J_{i-1}\big(q(b-u^*,a),g'\big) \\ & = J_i(b,g). \end{aligned} \tag{A.4}$$

We have proved that if $J_{i-1}(b,g)$ is increasing in b for all g then the same is true for $J_i(b,g)$. Therefore, by induction, $J^*_\alpha(b,g) = \lim_{i\to\infty} J_i(b,g)$ is increasing in b for all g.

□

Lemma A.0.2. *For all $0 \le b_1 < b_2 \le B$ and all $0 < g < K$, $J^*_\alpha(b_2,g) - J^*_\alpha(b_1,g)$ is upper bounded when β increases.*

Proof. Let u_1^* be the optimal transmission rate in state (b_1, g) then

$$J_\alpha^*(b_1, g) = C_I(b_1, g, u_1^*) + C_F(b_1, g, u_1^*). \tag{A.5}$$

Now let $u_2 = u_1^* + b_2 - b_1$, then as $J_\alpha^*(b_2, g)$ is the optimal cost associated with state (b_2, g), we must have:

$$J_\alpha^*(b_2, g) \leq C_I(b_2, g, u_2) + C_F(b_2, g, u_2). \tag{A.6}$$

Therefore

$$\begin{aligned}
&J_\alpha^*(b_2, g) - J_\alpha^*(b_1, g) \\
&\leq C_I(b_2, g, u_2) + C_F(b_2, g, u_2) - C_I(b_1, g, u_1^*) + C_F(b_1, g, u_1^*) \qquad \text{(A.7)} \\
&= C_I(b_2, g, u_2) - C_I(b_1, g, u_1^*) + C_F(b_2, g, u_2) - C_F(b_1, g, u_1^*).
\end{aligned}$$

As $u_2 = u_1^* + b_2 - b_1$, we have $C_F(b_2, g, u_2) = C_F(b_1, g, u_1^*)$ while

$$C_I(b_2, g, u_2) - C_I(b_1, g, u_1^*) = P(u_1^* + b_2 - b_1, g, \overline{P}_b) - P(u_1^*, g, \overline{P}_b).$$

Therefore

$$J_\alpha^*(b_2, g) - J_\alpha^*(b_1, g) \leq \frac{WN_o}{\gamma_g}\big(f(u_1^* + b_2 - b_1, \overline{P_b}) - f(u_1^*, \overline{P_b})\big). \tag{A.8}$$

It is clear that the left hand side of (A.8) is bounded when β increases, so the proof is completed. □

Note that Lemma A.0.2 is for situation in which the channel state $g > 0$, when $g = 0$, we have the following lemma.

Lemma A.0.3. *For all* $0 \leq b_1 < b_2 \leq B$, $J_\alpha^*(b_2, 0) - J_\alpha^*(b_1, 0)$ *increases without bound when* β *increases.*

Proof. When the channel is in state 0, no transmission is possible, therefore

$$\begin{aligned}
J^*_\alpha(b_1,0) &= C_I(b_1,0,0) + \alpha \sum_{g=0}^{K-1}\sum_{a=0}^{\infty} P_G(0,g)p_A(a)J^*_\alpha\big(q(b_1,a),g\big),\\
J^*_\alpha(b_2,0) &= C_I(b_2,0,0) + \alpha \sum_{g=0}^{K-1}\sum_{a=0}^{\infty} P_G(0,g)p_A(a)J^*_\alpha\big(q(b_2,a),g\big).
\end{aligned} \tag{A.9}$$

Therefore,

$$\begin{aligned}
&J^*_\alpha(b_2,0) - J^*_\alpha(b_1,0)\\
&= C_I(b_2,0,0) - C_I(b_1,0,0)\\
&\quad + \alpha \sum_{g=0}^{K-1}\sum_{a=0}^{\infty} P_G(0,g)p_A(a)\Big(J^*_\alpha\big(q(b_2,a),g\big) - J^*_\alpha\big(q(b_1,a),g\big)\Big) \qquad (A.10)\\
&> C_I(b_2,0,0) - C_I(b_1,0,0)\\
&= \beta \times \big(L(b_2,0) - L(b_1,0)\big).
\end{aligned}$$

The inequality in (A.10) is due to Lemma A.0.1. From (A.10), it is clear that $J^*_\alpha(b_2,0) - J^*_\alpha(b_1,0)$ increases without bound when β increases and the proof is completed. □

Using the results of Lemmas A.0.1, A.0.2, and A.0.3, let us prove the Lemma 3.3.1.

Lemma 3.3.1. *For each buffer state $b > 1$, there exists a constant β_o such that for every $\beta > \beta_o$ and $0 \le u_1 < u_2 \le b$, the following inequality holds:*

$$\Delta_I(b,1,u_1,u_2) - \Delta_I(b,2,u_1,u_2) < \Delta_F(b,1,u_1,u_2) - \Delta_F(b,2,u_1,u_2). \tag{A.11}$$

Proof. First of all, we have

$$\begin{aligned}
&\Delta_I(b,1,u_1,u_2) - \Delta_I(b,2,u_1,u_2)\\
&\qquad = WN_o\left(f(u_2,\overline{P_b}) - f(u_1,\overline{P_b})\right)\left(\frac{1}{\gamma_1} - \frac{1}{\gamma_2}\right).
\end{aligned} \tag{A.12}$$

Therefore, the left hand side of (A.11) does not depend on β. For the right hand side of (A.11), we have:

$$\Delta_F(b,g,u_1,u_2) = \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} P_G(g,g') p_A(a) (J^*_\alpha(q(b-u_1,a),g') - J^*_\alpha(q(b-u_2,a),g')). \tag{A.13}$$

Now

$$\begin{aligned} &\Delta_F(b,1,u_1,u_2) - \Delta_F(b,2,u_1,u_2) \\ &= \alpha \sum_{g'=0}^{K-1} \sum_{a=0}^{\infty} \Big[\big(P_G(1,g') - P_G(2,g')\big) \times p_A(a) \\ &\times \Big(J^*_\alpha\big(q(b-u_1,a),g'\big) - J^*_\alpha\big(q(b-u_2,a),g'\big)\Big) \Big] \end{aligned} \tag{A.14}$$

$$\begin{aligned} &= \alpha \sum_{g'=1}^{K-1} \sum_{a=0}^{\infty} \Big[\big(P_G(1,g') - P_G(2,g')\big) \times p_A(a) \\ &\times \Big(J^*_\alpha\big(q(b-u_1,a),g'\big) - J^*_\alpha\big(q(b-u_2,a),g'\big)\Big) \Big] \\ &+ \alpha \sum_{a=0}^{\infty} \Big[\big(P_G(1,0) - P_G(2,0)\big) \times p_A(a) \\ &\times \Big(J^*_\alpha\big(q(b-u_1,a),0\big) - J^*_\alpha\big(q(b-u_2,a),0\big)\Big) \Big]. \end{aligned} \tag{A.15}$$

When β increases, from Lemmas A.0.1 and A.0.2, the first term in (A.15) is always lower bounded while from Lemma A.0.3, the second term increases without bound. This combined with (A.12) completes the proof. □

APPENDIX B

PROOF OF LEMMA 5.3.2

Lemma 5.3.2. *For any stationary feasible adaptive scheduling/transmission policy $\phi \in \mathbf{\Psi}^{st}$, let L_o^{ϕ} be the total packet loss rate of all users and P_n^{ϕ} be the average power consumed by user n when ϕ is employed, there exists a non-stationary policy $\psi \in \mathbf{\Psi}$ such that*

$$L_o^{\psi} \; = \; L_o^{\phi} \quad \textit{while} \quad P_m^{\psi} \; = \; \frac{1}{N}\sum_{n=1}^{N} P_n^{\phi}, \quad \forall m \in \mathcal{N}, \tag{B.1}$$

where L_o^{ψ} is the total packet loss rate and P_m^{ψ} is the average power consumed by user m when policy ψ is employed.

Proof. Given a stationary policy $\phi \in \mathbf{\Psi}^{st}$, we will construct a non-stationary policy $\psi \in \mathbf{\Phi}$ that satisfies (5.7). This is done by first formulating $N-1$ other stationary policies, $\phi^1, \phi^2, \ldots \phi^{N-1}$, and then time sharing $\phi, \phi^1, \phi^2, \ldots \phi^{N-1}$.

Note again that each stationary scheduling/transmission policy ϕ^k, $k = 1, 2, \ldots N-1$, is completely specified by the vector of transmission rates assigned to the N users in each system state. In time slot i, the system state is $\boldsymbol{S}_i = (B_i^1, B_i^2, \ldots B_i^N, \; G_i^1, G_i^2, \; \ldots G_i^N)$. Let $\boldsymbol{I}^k$, $\boldsymbol{B}_i^k$, and $\boldsymbol{G}_i^k$ be the N-element vectors obtained after carrying out k right cyclic shifts on vectors $(1, 2, \ldots N)$, $(B_i^1, B_i^2, \ldots B_i^N)$, and $(G_i^1, G_i^2, \ldots G_i^N)$ respectively. We then set

$$\phi^k(\boldsymbol{S}_i, n) = \phi(\boldsymbol{S}_i^k, \boldsymbol{I}^k(n)) \tag{B.2}$$

where $\boldsymbol{S}_i^k = (\boldsymbol{B}_i^k, \boldsymbol{G}_i^k)$ and $\boldsymbol{I}^k(n)$ is the n^{th} element of $\boldsymbol{I}^k$. Note that policy ϕ^k is nothing but policy ϕ being applied to a modified system in which the sequence of N users is permutated by carrying out k right cyclic shifts. As all N users in the system are symmetric, i.e., they have i.i.d. data arrival processes, i.i.d.

channel processes, and the same buffer lengths, when ϕ^k is employed the total packet loss rate and average transmit powers are

$$L_o^{\phi^k} \;=\; L_o^{\phi}, \quad \text{while} \quad P_n^{\phi^k} = \boldsymbol{P}^k(n) \tag{B.3}$$

where $\boldsymbol{P}^k$ is the vector obtained after k right cyclic shifts of $(P_1^{\phi}, P_2^{\phi}, \ldots P_N^{\phi})$.

For convenient of notation, we let $\phi^0 = \phi$. Now, based on N stationary policies $\phi^0, \phi^1, \phi^2, \ldots \phi^{N-1}$, we construct a non-stationary adaptive scheduling/transmission policy $\psi = \{\phi_i\}$ that satisfies

$$\phi_i = \phi^k, \text{ where } k = \left\lfloor \frac{\bmod (i, Nt)}{t} \right\rfloor, \quad \forall i = 0, 1, 2, \ldots. \tag{B.4}$$

Note that in (B.4), $\bmod (x, y)$ gives the remainder on division of x by y while $\lfloor x \rfloor$ is the floor function. To put it simple, the system is control on a frame-by-frame basis, each frame is of length Nt time slots. During each frame, N policies $\phi^0, \phi^1, \phi^2, \ldots \phi^{N-1}$ are employed sequentially, each for t consecutive time slots. When $t \to \infty$ the averaging effect takes place and during each frame, we have the average total packet loss rate is L_o^{ϕ} while the average power consumed by user m is

$$\frac{1}{N} \sum_{k=0}^{N-1} P_m^{\phi^k} = \frac{1}{N} \sum_{n=1}^{N} P_n^{\phi}.$$

This shows that ψ satisfies (B.1). $\square$

APPENDIX C

PUBLICATION LIST

C.1 International Conferences

1. A. T. Hoang and M. Motani. Exploiting Wireless Broadcast in Spatially-Correlated Sensor Networks. In *Proceedings 2005 IEEE International Conference on Communications (ICC)*, May 2005.

2. A. T. Hoang and M. Motani. Collaborative broadcasting and compression in cluster-based wireless sensor networks. In *Proceedings 2nd European Workshop on Wireless Sensor Networks (EWSN)*, Jan. 2005.

3. A. T. Hoang and M. Motani. Decoupling multiuser cross-layer adaptive transmission. In *Proceedings 2004 IEEE International Conference on Communications (ICC)*, June 2004, pp. 3061-3065.

4. A. T. Hoang and M. Motani. Buffer and channel adaptive transmission over fading channels with imperfect channel state information. In *Proceedings 2004 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2004, pp. 1891-1896.

5. A. T. Hoang and M. Motani. Adaptive sensing and transmitting for increasing lifetime of energy constrained sensor networks. In *Proceedings 38th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2004, Princeton University.

6. M. Motani and A. T. Hoang. An instance of multiuser diversity in wireless networks. In *Proceedings 2003 IEEE International Symposium on Information Theory (ISIT)*, July 2003, p. 443.

7. A. T. Hoang and M. Motani. Buffer and channel adaptive modulation for transmission over fading channels. In *Proceedings 2003 IEEE International Conference on Communications (ICC)*, July 2003, pp. 2748-2752.

8. A. T. Hoang and M. Motani. Buffer control using adaptive MQAM for wireless channels". In *Proceedings IFIP TC6/WG6.8 Working Conference on Personal Wireless Communications (PWC)*, Oct. 2002, pp. 77-104.

C.2 Book Chapter

1. A. T. Hoang and M. Motani. Adaptive sensing and reporting in energy constrained sensor networks. *Sensor Network Operations*, Wiley-IEEE Press, May 2006.

C.3 Journals

1. A. T. Hoang and M. Motani. Cross-layer adaptive transmission: Optimal strategies in fading channels. *Under review for IEEE Transactions on Communications.*

2. A. T. Hoang and M. Motani. Cross-layer adaptive transmission: Coping with incomplete system state information. *Submitted to IEEE Transactions on Wireless Communications.*

3. A. T. Hoang and M. Motani. Collaborative broadcasting and compression in cluster-based wireless sensor networks. *Under review for ACM Transactions on Sensor Networks.*

4. A. T. Hoang and M. Motani. Adaptive Scheduling and Transmission for multiple access wireless channels. *In preparation.*

BIBLIOGRAPHY

[AKR+01] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. Providing quality of service over a shared wireless link. *IEEE Communications Magazine*, 39(2):150–154, Feb. 2001.

[AKR+04] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, and P. Whiting. Scheduling in a queueing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences*, 18:191–217, 2004.

[AN98] A. T. Andersen and B. F. Nielsen. A markovian approach for modeling packet traffic with long-range dependence. *IEEE Journal on Selected Areas in Communications*, 16(5):719–732, 1998.

[ANJ05] S. Agnihotri, P. Nuggehalli, and H. S. Jamadagni. On maximizing lifetime of a sensor network. In *Proceedings of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Taormina/Giardini Naxos, June 2005.

[ASSC02] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102–114, August 2002.

[BG02] R. A. Berry and R. G. Gallager. Communication over fading channels with delay constraints. *IEEE Transactions Information Theory*, 48(5):1135–1149, May 2002.

[BTD01] S. Basagni, D. Turgut, and S. K. Das. Mobility-adaptive protocol for managing large ad hoc networks. In *Proceedings of IEEE International Conference on Communications (ICC)*, pages 1539–1543, Jun. 2001.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 1st edition, 2004.

[BW01] S. Borst and P. Whiting. Dynamic rate control algorithms for HDR throughput optimization. In *Proceedings of IEEE INFOCOM 2001*, pages 976–985, Anchorage, Alaska, Apr. 2001.

[CC99] B. Collins and R. Cruz. Transmission policy for time varying channel with average delay constraints. In *Proceedings of 1999 Allerton Conf. on Commun. Control and Comp*, pages 1–9, Monticello, IL, 1999.

[CG05] X. Cai and G. B. Giannakis. Adaptive psma accounting for channel estimation and prediction errors. *IEEE Transactions Wireless Communications*, 4(1):246–256, Jan. 2005.

[CLV04] R. Cristescu, B. Lozano, and M. Vetterli. On network correlated data gathering. In *Proceedings of IEEE INFOCOM 2004*, pages 2571–2582, Hong Kong, March 2004.

[CPR03] J. Chou, D. Petrovic, and K. Ramchandran. A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks. In *Proceedings of IEEE INFOCOM 2003*, pages 1054–1062, San Francisco, March 2003.

[CWKS97] B. P. Crow, I. Widjaja, J. G. Kim, and P. T. Sakai. IEEE 802.11 wireless local area networks. *IEEE Communications Magazine*, 35(9):116–126, Sep. 1997.

[DG97] T. Dean and R. Givan. Model minimization in Markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI, 1997*, pages 106–111, 1997.

[DGM+04] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *Proceedings of 30th International Conference on Very Large Data Bases (VLDB)*, pages 588–599, August 2004.

[DMS+03] K. Das, J. Marks, E. Sharkawi, P. Arabshahi, and A. Gray. Minimum power broadcast trees for wireless networks: Integer programming formations. In *Proceeding of IEEE INFOCOM'03*, San Francisco, March 2003.

[FS83] G. J. Foschini and J. Salz. Digital communications over fading radio channels. *Bell Syst. Tech. J.*, pages 429–456, February 1983.

[Gal68] R. G. Gallager. *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[GC97] A. J. Goldsmith and S. G. Chua. Variable-rate variable-power mqam for fading channels. *IEEE Transactions on Communications*, 45(10):1218–1230, October 1997.

[GE03] A. Goel and D. Estrin. Simultaneous optimization for concave cost: Single sink aggregation or single source buy-at-bulk. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pages 499–505, 2003.

[GKS03] M. Goyal, A. Kumar, and V. Sharma. Power constrained and delay optimal policies for scheduling transmission over a fading channel. In *Proceedings of IEEE INFOCOM'03*, pages 311–320, San Francisco, March 2003.

[Goe99] D. L. Goeckel. Adaptive coding for time-varying channel using outdated fading estimates. *IEEE Transactions Communications*, 47(6):844–855, Jun. 1999.

[Gud91] M. Gudmundon. Correlation model for shadow fading in mobile radio systems. *Electronics Letters*, 27(23):2145–2146, 1991.

[GV97] A. J. Goldsmith and P. P. Varaiya. Capacity of fading channels with channel side information. *IEEE Transactions Information Theory*, 43(6):1986–1992, November 1997.

[GW02] A. J. Goldsmith and S. B. Wicker. Design challenges for energy-constrained ad hoc wireless networks. *IEEE Wireless Communications Magazine*, 9(4):8–27, August 2002.

[Haa00] J. C. Haartsen. The bluetooth radio system. *IEEE Personal Communications*, 7(1):28–36, 2000.

[HCB00] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan. Energy-efficient communication protocol for wireless sensor networks. In *Proceedings of the 33th Hawaii International Conference on System Sciences (HICSS-33)*, pages 1–10. IEEE Computer Society, January 2000.

[HCB02] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan. An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions Wireless Communications*, 1(4):660–670, October 2002.

[HGG02] T. Holliday, A. J. Goldsmith, and P. Glynn. Wireless link adaptation policies: Qos for deadline constrained traffic with imperfect channel estimates. In *Proceedings of IEEE International Conference on Communications (ICC)*, New York, Apr. 2002.

[HL95] F. Hiller and G. Lieberman. *Introduction to Operation Research.* McGraw-Hill Inc, 1995.

[HM02] A. T. Hoang and M. Motani. Buffer control using adaptive mqam for wireless channels. In *Proceedings of IFIP TC6/WG6.8 Working Conference on Personal Wireless Communications (PWC 2002)*, pages 97–104, Singapore, October 2002.

[HM03] A. T. Hoang and M. Motani. Buffer and channel adaptive modulation for transmission over fading channels. In *Proceedings of IEEE International Conference on Communications (ICC 2003)*, pages 2748–2752, Anchorage, Alaska, July 2003.

[HM04a] A. T. Hoang and M. Motani. Buffer and channel adaptive transmission over fading channels with imperfect channel state information. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC 2004)*, pages 1891–1896, Atlanta, Mar. 2004.

[HM04b] A. T. Hoang and M. Motani. Decoupling multiuser cross-layer adaptive transmission. In *Proceedings of IEEE International Conference on Communications (ICC 2004)*, pages 3061–3065, Paris, Jun. 2004.

[HM05a] A. T. Hoang and M. Motani. Collaborative broadcasting and compression in cluster-based wireless sensor networks. In *Proceedings 2nd European Workshop on Wireless Sensor Networks (EWSN 2005)*, pages 197–206, Istanbul, January 2005.

[HM05b] A. T. Hoang and M. Motani. Collaborative broadcasting and compression in cluster-based wireless sensor networks. *Under review for ACM Transactions on Sensor Networks*, 2005.

[HM05c] A. T. Hoang and M. Motani. Cross-layer adaptive transmission: Coping with incomplete system state information. *Submitted to IEEE Transactions on Wireless Communications*, 2005.

[HM05d] A. T. Hoang and M. Motani. Cross-layer adaptive transmission: Optimal strategies in fading channels. *Under review for IEEE Transactions on Communications*, 2005.

[HM05e] A. T. Hoang and M. Motani. Exploiting wireless broadcast in spatially correlated sensor networks. In *Proceedings of IEEE International Conference on Communications (ICC 2005)*, Seoul, Korea, May 2005.

[IGE00] C. Intanagonwiwat, R. Govindan, and D. Estrin. Directed diffusion: A scalable and robust communication paradigm for sensor networks. In *Proceedings of ACM MobiCom'00*, Boston, Aug. 2000.

[Jak74] W. C. Jakes. *Microwave Mobile Communications*. John Willey and Sons, 1974.

[JP04] A. Jindal and K. Psounis. Modeling spatially-correlated sensor network data. In *Proceedings of IEEE International Conference on Sensor and Ad hoc Communications and Networks (SECON) 2004*, pages 162–171, 2004.

[KDN03] K. Kalpakis, K. Dasgupta, and P. Namjoshi. Efficient algorithms for maximum lifetime data gathering and aggregation in wireless sensor networks. *ACM Computer Networks*, 42(6):697–716, Aug. 2003.

[KEW02] B. Krishnamachari, D. Estrin, and S. Wicker. Modeling data centric routing in wireless sensor networks. In *USC Computer Engineering Technical Report CENG*, pages 12–44, 2002.

[KH95] R. Knopp and P. Humblet. Information capacity and power control in single cell multiuser communications. In *Proceedings of IEEE International Conference on Communications (ICC'95)*, pages 331–335, Seattle, June 1995.

[KK05] V. Kawadia and P. R. Kumar. A cautionary perspective on cross layer design. *IEEE Wireless Communication Magazine*, 12(1):3–11, 2005.

[KKP99] J. M. Kahn, R. H. Katz, and K. S. J. Pister. Next century challenges: Mobile networking for smart dust. In *Proceedings ACM MobiCom 99*, pages 271–278, 1999.

[KQ03] K. Kumaran and L. Qian. Uplink scheduling in CDMA packet-data systems. In *Proceedings of IEEE INFOCOM 2003*, San Francisco, Mar. 2003.

[KV86] P. R. Kumar and P. Varaiya. *Stochastic Systems: estimation, identification, and adaptive control.* Englewood Cliffs, N.J.: Prentice Hall, 1986.

[LBH03] P. Liu, R. Berry, and M. L. Honig. Delay-sensitive packet scheduling in wireless networks. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, New Orleans, Mar. 2003.

[LCK95] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In *Proceedings of Twelfth International Conference on Machine Learning*, pages 362–370, 1995.

[LCS01] X. Liu, E. K. P. Chong, and N. B. Shroff. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. *IEEE Journal of Selected Areas in Communications*, 19(10):2053–2065, Oct. 2001.

[LK03] Y. Liu and E. Knightly. Opportunistic fair scheduling over multiple wireless channels. In *Proceedings of IEEE INFOCOM 2003*, San Francisco, Mar. 2003.

[Lov91] W. S. Lovejoy. A survey of algorithmic methods for partially observable markov decision processes. *Annals of Operation Research*, 28(1):47–65, 1991.

[ML02] D. Melo and M. Liu. Analysis of energy consumption and lifetime of heterogeneous wireless sensor networks. In *Proceeding of IEEE Globecom 2002*, 2002.

[NMR03] M. J. Neely, E. Modiano, and C. E. Rohrs. Power allocation and routing in multibeam satellites with time-varying channels. *IEEE/ACM Transactions on Networking*, 11(1):138–152, 2003.

[NPB95] I. Nourbakhsh, R. Powers, and S. Birchfield. Dervish an office-navigating robot. *AI Magazine*, 16(2):53–60, 1995.

[OHH04] G. E. Oien, H. Holm, and K. J. Hole. Impact of channel prediction on adaptive coded modulation performance in rayleigh fading. *IEEE Transactions on Vehicular Technology*, 53(3):758–769, May 2004.

[PKG04] S. Pattem, B. Krishnamachari, and R. Govindan. The impact of spatial correlation on routing with compression in wireless sensor networks. In *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, pages 28–35, New York, April 2004. ACM Press.

[PR99] S. Pradhan and K. Ramchandran. Distributed source coding using syndromes: Design and construction. In *Proceedings of the Data Compression Conference (DCC)*, page 158, Washington, DC, USA, 1999. IEEE Computer Society.

[Pro01] J. G. Proakis. *Digital Communications*. McGraw-Hill International, 4th edition, 2001.

[QB04] X. Qin and R. Berry. Opportunistic splitting algorithms for wireless networks. In *Proceedings of IEEE INFOCOM 2004*, Hong Kong, Mar. 2004.

[Rap96] T. Rappaport. *Ad hoc Mobile Wireless Networks: Protocols and Systems*. Prentice Hall, 1996.

[RSA04] D. Rajan, A. Sabharwal, and B. Aazhang. Delay bounded packet scheduling of bursty traffic over wireless channels. *IEEE Transactions on Information Theory*, 50(1):125–144, 2004.

[RSPS02] V. Raghunathan, C. Schurgers, S. Park, and M. Srivastava. Energy aware wireless microsensor networks. *IEEE Signal Processing Magazine*, 19(2):40–50, March 2002.

[SCI+01] E. Shih, S. Cho, N. Ickes, R. Min, A. Sinha, A. Wang, and A. Chandrakasan. Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks. In *Proceedings of the 7th ACM International Conference on Mobile Computing and Networking (MobiCom 2001)*, pages 272–287. ACM Press, July 2001.

[SRB01] A. Sabharwal, D. Rajan, and B.Aazhang. Dual problems in power control. In *Proceedings of Allerton Conference on Communication, Control and Computing*, Monticello, IL, Oct. 2001.

[SRK03] S. Shakkotai, T. Rappaport, and P. Karlsson. Cross-layer design for wireless networks. *IEEE Communications Magazine*, 41(10):74–80, 2003.

[SS02a] A. Scaglione and S. Servetto. On the interdependence of routing and data compression in multihop-sensor networks. In *Proceedings of the 8th ACM International Conference on Mobile Computing and Networking (MobiCom 2002)*, pages 140–147, Atlanta, September 2002.

[SS02b] S. Shakkottai and A. L. Stolyar. Scheduling for multiple flows sharing a time-varying channel: The exponential rule. *Analytic Methods in Applied Probability. In Memory of Fridrih Karpelevich. Yu. M. Suhov, Editor*, 207:185–202, 2002.

[SSZ01] I. Stojmenovic, M. Seddigh, and J. Zunic. Internal nodes based broadcasting in wireless networks. In *Proceeding of the 34th Hawaii International Conference on System Sciences*, 2001.

[SW73] D. Slepian and K. Wolf. Noiseless encoding of correlated information sources. *IEEE Transactions on Information Theory*, IT-19(4):471–480, Jul. 1973.

[TE93] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions Information Theory*, 39(2):466–478, Mar. 1993.

[TH98a] D. Tse and S. Hanly. Multi-access fading channels: Part i: Polymatroid structure, optimal resource allocation and through-put capacities. *IEEE Transactions on Information Theory*, 44(7):2796–2815, 1998.

[TH98b] D. Tse and S. Hanly. Multi-access fading channels: Part ii: Delay-limited capacities. *IEEE Transactions on Information Theory*, 44(7):2816–2831, 1998.

[VAT03] P. Venkitasubramaniam, S. Adireddy, and L. Tong. Opportunistic aloha and cross-layer design for sensor networks. In *Proceedings of IEEE MILCOM*, Boston, Oct. 2003.

[Vin05] S. Vineet. Cross-layer design for communication systems. Master's thesis, National University of Singapore, 2005.

[Vuc91] B. Vucetic. An adaptive coding scheme for time-varying channels. *IEEE Transactions Communications*, 39(5):653–663, May 1991.

[WM95] H. S. Wang and N. Moayeri. Finite-state markov channel - a useful model for radio communication channels. *IEEE Transactions on Vehicular Technology*, 44(1):163–171, February 1995.

[WNE00] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides. On the construction and energy-efficient broadcast and multicast trees in wireless networks. In *Proceedings of IEEE INFOCOM'00*, Tel-Aviv, Mar. 2000.

[WS95] W. T. Webb and R. Steele. Variable rate qam for mobile radio. *IEEE Transactions on Communications*, 43(7):2223–2230, July 1995.

[WZ76] A. D. Wyner and J. Ziv. The rate-distortion function for source coding with side information at decoder. *IEEE Trans. on Inform. Theory*, IT-22(1):1–10, January 1976.

[ZW02] D. Zhang and K. Wasserman. Transmission schemes for time-varying wireless channels with partial state observations. In *Proceedings of IEEE INFOCOM'02*, pages 467–476, New York, Jun. 2002.